A Smart Data Ecosystem for the Monitoring of Financial Market Irregularities

Lewis Evans

PhD 2022

A Smart Data Ecosystem for the Monitoring of Financial Market Irregularities

LEWIS EVANS

A thesis submitted in partial fulfilment of the requirements of Manchester Metropolitan University for the degree of Doctor of Philosophy

Department of Computing and Mathematics
Manchester Metropolitan University

2022

Declaration of Authorship

I, Lewis EVANS, declare that this thesis titled, “A Smart Data Ecosystem for the Monitoring of Financial Market Irregularities” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- This thesis does not exceed the regulation length of 80 000 words.

Signed:

Date: 4/3/2022

Abstract

Investments made on the stock market depend on timely and credible information being made available to investors. Such information can be sourced from online news articles, broker agencies, and discussion platforms such as financial discussion boards and Twitter. The monitoring of such discussion is a challenging yet necessary task to support the transparency of the financial market. Although financial discussion boards are typically monitored by administrators who respond to other users reporting posts for misconduct, actively monitoring social media such as Twitter remains a difficult task.

Users sharing news about stock-listed companies on Twitter can embed cashtags in their tweets that mimic a company's stock ticker symbol (e.g. TSCO on the London Stock Exchange refers to Tesco PLC). A cashtag is simply the ticker characters prefixed with a '$' symbol, which then becomes a clickable hyperlink – similar to a hashtag. Twitter, however, does not distinguish between companies with identical ticker symbols that belong to different exchanges. TSCO, for example, refers to Tesco PLC on the London Stock Exchange but also refers to the Tractor Supply Company listed on the NASDAQ. This research has referred to such scenarios as a 'cashtag collision'. Investors who wish to capitalise on the fast dissemination that Twitter provides may become susceptible to tweets containing colliding cashtags. Further exacerbating this issue is the presence of tweets referring to cryptocurrencies, which also feature cashtags that could be identical to the cashtags used for stock-listed companies. A system that is capable of identifying stock-specific tweets by resolving such collisions, and assessing the credibility of such messages, would be of great benefit to a financial market monitoring system by filtering out non-significant messages. This project has involved the design and development of a novel, multi-layered, smart data ecosystem to monitor potential irregularities within the financial market. This ecosystem is primarily concerned with the behaviour of participants' communicative practices on discussion platforms and the activity surrounding company events (e.g. a broker rating being issued for a company). A wide array of data sources – such as tweets, discussion board posts, broker ratings, and share prices – is collected to support this process. A novel data fusion model fuses together these data sources to provide synchronicity to the data and allow easier analysis of the data to be undertaken by combining data sources for a given time window (based on the company the data refers to and the date and time). This data fusion model, located within the data layer of the ecosystem, utilises supervised machine learning classifiers - due to the domain expertise needed to accurately describe the origin of a tweet in a binary way - that are trained on a novel set of features to classify tweets as being related to a London Stock Exchange-listed company or not. Experiments involving the training of such classifiers have achieved accuracy scores of up to 94.9%.

The ecosystem also adopts supervised learning to classify tweets concerning their credibility. Credibility classifiers are trained on both general features found in all tweets, and a novel set of features only found within financial stock tweets. The experiments in which these credibility classifiers were trained have yielded AUC scores of up to 94.3.

Once the data has been fused, and irrelevant tweets have been identified, unsupervised clustering algorithms are then used within the detection layer of the ecosystem to cluster tweets and posts for a specific time window or event as potentially irregular. The results are then presented to the user within the presentation and decision layer, where the user may wish to perform further analysis or additional clustering.

Publications

The work presented in this thesis has been undertaken during my candidature at Manchester Metropolitan University and report on various aspects of the work undertaken.

Journals

- **Evans, L.**, Owda, M., Crockett, K., & Fernandez Vilas, A. (2021). Credibility assessment of financial stock tweets. *Expert Systems with Applications*, 168, 114351.
- **Evans, L.**, Owda, M., Crockett, K., & Vilas, A. F. (2019). A methodology for the resolution of cashtag collisions on Twitter – A natural language processing & data fusion approach. *Expert Systems with Applications*, 127, 353–369.
- Vilas, A. F., Díaz Redondo, R. P., Crockett, K., Owda, M., & **Evans, L.** (2019). Twitter permeability to financial events: an experiment towards a model for sensing irregularities. *Multimedia Tools and Applications*, 78(7).

Conferences

- **Evans, L.**, Owda, M., Crockett, K., & Vilas, A. F. (2018). Big Data Fusion Model for Heterogeneous Financial Market Data (FinDF). *Proceedings of the 2018 Intelligent Systems Conference (IntelliSys)*, 1085–1101.
- Vilas, A. F., **Evans, L.**, Owda, M., Díaz Redondo, R. P., & Crockett, K. (2017). Experiment for analysing the impact of financial events on twitter. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10393 LNCS, 407–419.

Acknowledgements

I would first like to acknowledge the financial support that was given to me by the department which provided funding through my role as a Post Graduate Teaching Assistant.

Countless people have supported me throughout this journey. I will forever be indebted to my supervisors, **Prof. Keeley Crockett**, **Dr. Majdi Owda**, and **Prof. Ana Férnandez Vilas**. My supervisors have been stalwart in their support throughout the entirety of this project and have lifted my spirits more times than I care to admit.

I would also like to extend my appreciation to the **Financial Conduct Authority** for allowing me to present some of my work at their London headquarters. The comments I received were incredibly useful and have helped hone certain aspects of my research. I would also like to thank the five financial market experts that helped evaluate the ecosystem developed as part of this research - their feedback has been invaluable.

My position as a PGTA has afforded me invaluable teaching opportunities. It would be remiss of me to not mention the countless students who have helped shaped my journey. I am grateful to each and every one of them and hope that they have learned as much from me as I did from them. I would also like to extend my appreciation to the Head of Department, **Prof. Darren Dancey**, and the unit leader for the module I delivered, **Dr. David McLean**, for providing me with such rewarding teaching opportunities throughout my role.

As with any significant undertaking, it would not have been possible without the support of friends. I have been fortunate to forge friendships with a legion of people during my time at MMU, and have made many fond memories. A special mention should be made to **Naomi Adel**, **Dr. Matthew Crossley**, **Dr. Kristopher Welsh**, , **Zoe Bartlett**, and **Dr. Amy Khalfay**. A special thanks must also go to **Dr. Elaine Duffin** who kindly agreed to proof read this thesis towards its later stages. Last but by no means least, **Lauren Smart** and **Alasdair Ward** have kept me sane throughout this PhD and provided endless laughter and support.

Contents

Declaration of Authorship iii

Abstract v

Publications vii

Acknowledgements ix

1 Introduction 1
- 1.1 Overview 1
- 1.2 Motivation 1
  - 1.2.1 Efficient Market Hypothesis 2
- 1.3 Research Aim 2
- 1.4 Research Objectives 3
- 1.5 Research Questions 3
- 1.6 Contributions 4
- 1.7 Thesis Overview 5

2 Background: Financial Market Landscape 7
- 2.1 Overview 7
- 2.2 A Brief History of Financial Markets 7
- 2.3 London Stock Exchange 8
  - 2.3.1 Main Market 8
  - 2.3.2 Alternative Investment Market 9
- 2.4 Stock Market Participants 9
  - 2.4.1 Companies 10
  - 2.4.2 Investors 11
  - 2.4.3 Broker Agencies 11
  - 2.4.4 Banks 11
  - 2.4.5 Regulator: Financial Conduct Authority 12
- 2.5 Stock Market Information Sources 13
  - 2.5.1 Financial Discussion Boards (e.g. London South East) 15
  - 2.5.2 Social Media (e.g. Twitter) 16
  - 2.5.3 Broker Ratings (via Broker Agencies) 17
  - 2.5.4 Regulatory News Services (RNS) 17
- 2.6 Chapter Summary 18

3 Background: Detection of Irregularities 19
- 3.1 Overview 19
- 3.2 Defining an Irregularity 19
- 3.3 Documented Financial Market Irregularities 20
- 3.4 Irregularity Detection in Financial Markets 21
  - 3.4.1 Irregularity Detection Challenges 22

3.5 Financial Market Monitoring 23
3.5.1 Proposed FMMS Frameworks 23
3.5.2 Models Supporting Monitoring Systems 29
3.6 Chapter Summary 30

4 Research Methodology 31
4.1 Overview 31
4.2 Smart Data Ecosystem 31
4.2.1 What is a 'Smart' Data Ecosystem? 31
4.2.2 Smart Data Ecosystem Overview 33
4.2.3 Data Layer 33
4.2.4 Detection Layer 34
4.2.5 Presentation & Decision Layer 34
4.2.6 SDE Companies 35
4.3 Chapter Summary 35

5 Data Layer: Foundations & Data Fusion Model 37
5.1 Overview 37
5.2 Data Fusion 38
5.3 Existing Data Fusion Models 38
5.3.1 Joint Directors of Laboratories (JDL) Model 39
5.3.2 Dasarathy Fusion Model 40
5.4 Data Fusion Challenges 42
5.5 Data Fusion in the Stock Market 43
5.5.1 Stock price prediction through data fusion 43
5.5.2 Risk/return forecasting 45
5.5.3 Summary of related work 46
5.6 Data Layer: Data Feeders & Collection 46
5.6.1 Tweet Collection 48
5.6.2 London South East Data Collection 48
5.6.3 Share Price Collection 50
5.7 Data Layer: Data Fusion Model 50
5.8 Chapter Summary 52

6 Data Layer: Resolving Colliding Cashtags within Tweets 53
6.1 Overview 53
6.2 Colliding Cashtags 54
6.2.1 Challenges associated with Colliding Cashtags 56
6.3 Related work utilising Cashtags 57
6.3.1 Disambiguation on Twitter 57
6.3.2 Stock Prediction using Cashtag Tweets 58
6.4 Experiment Overview 58
6.4.1 Phase 1: Experiment Preparation 58
6.4.2 Phase 2: Data Collection 59
6.4.3 Phase 3: Data Fusion & Corpora Creation 60
6.4.4 Phase 4: Machine Learning 60
6.5 Company Corpora Creation Methodology 62
6.5.1 Corpora Creation 62
6.6 Cashtag Collision Resolution Methodology (CCRM) 64
6.6.1 Annotated tweet dataset 65
6.6.2 Step 1: Creation of tweet sparse vector 66

6.6.3 Step 2: Company Corpora Consultation 66
6.6.4 Step 3: Feature Generation 66
6.6.5 Step 4: Classifier Training 67
6.6.6 Step 5: Performance Evaluation 67
6.7 Results & Discussion 68
6.7.1 Classifier Evaluation 68
6.7.2 Classifier Results 69
6.7.3 Logistic Regression 71
6.7.4 k-Nearest Neighbours 71
6.7.5 Decision Tree 71
6.7.6 Random Forest 72
6.7.7 Support Vector Machine 72
6.7.8 Discussion 73
6.8 Chapter Summary 73

7 Data Layer: Assessing Tweet Credibility 75
7.1 Overview 75
7.2 Related work on Credibility 76
7.2.1 Summary of Related Work 79
7.3 Credibility Assessment Methodology 80
7.3.1 Stage 1 - Data Collection 80
7.3.2 Stage 2 - Model Preparation 81
7.3.3 Stage 3 - Model Training 82
7.4 Feature Groups 82
7.4.1 General Features 83
7.4.2 Financial Features 84
7.5 Feature Selection 87
7.5.1 Filter Methods 87
7.5.2 Wrapper Methods 88
7.5.3 Embedded Methods 88
7.6 Experimental Design 89
7.6.1 Company Selection 89
7.6.2 Data Collection 89
7.6.3 Tweet Annotation 89
7.6.4 Assessing Feature Importance 91
7.7 Results & Discussion 94
7.8 Chapter Summary 96

8 Detection Layer: Detection of Financial Market Irregularities 99
8.1 Overview 99
8.2 Irregularity Detection via clustering 100
8.3 Event Generation Methodology 101
8.4 Irregularity Detection Methodology 101
8.4.1 Choosing the optimal number of clusters 103
8.4.2 Visualising the results via Principal Component Analysis (PCA) 104
8.5 Irregular Event Detection 105
8.6 Irregular Tweet Detection 106
8.7 Irregular FDB Post Detection 108
8.8 Chapter Summary 109

9 Ecosystem Evaluation 111
- 9.1 Overview 111
- 9.2 Evaluation Methodology 111
  - 9.2.1 Participant Identification and Shortlisting 112
  - 9.2.2 Ecosystem Scenarios 112
- 9.3 Interview Overview 115
- 9.4 Participant Overview 116
- 9.5 Evaluation Scenarios 116
  - 9.5.1 Scenario 1 - Resolving Cashtag Collisions between Companies . 117
  - 9.5.2 Scenario 2 - Filtering Noisy Cryptocurrency Tweets 119
  - 9.5.3 Scenario 3 - Financial Stock Tweet Credibility 121
  - 9.5.4 Scenario 4 - Clustering of Events 124
  - 9.5.5 Scenario 5 - Clustering of Event Tweets 126
  - 9.5.6 Scenario 6 - Clustering of Event FDB Posts 129
  - 9.5.7 Closing Remarks 131
- 9.6 Chapter Summary 132

10 Conclusions and Future Work 133
- 10.1 Overview 133
- 10.2 Review of Contributions 133
  - 10.2.1 Smart Data Ecosystem 133
  - 10.2.2 Data Fusion Model 134
  - 10.2.3 Resolving Cashtag Collisions 134
  - 10.2.4 Assessing the Credibility of Financial Stock Tweets 135
  - 10.2.5 Detection of potential irregularities 135
- 10.3 Review of Research Questions 135
- 10.4 Future Work 137
  - 10.4.1 Large-scale Tweet Collection 138
  - 10.4.2 Automatic Detection of Irregularities 138
  - 10.4.3 Additional Events 138

A Data Fusion Paper 141

B SDE Companies 151

C Cashtag Collision Paper 157

D Credibility Assessment of Financial Stock Tweets Paper 175

E Cashtag Collision Experiment Companies 193

F Credibility Classifier Features 197

G Detection Layer Clustering Features 201
- G.1 Event Clustering Features 202

H Participant Presentation 205

I Participant Information Sheet 217

J System Evaluation Ethical Approval 223

List of Figures

1.1 Links between the research objectives, thesis chapters, and research outputs 5

2.1 Stock Market Participants 10
2.2 Fines issued by the FCA (2013 - 2020) (Financial Conduct Authority 2021) 13
2.3 Financial data sources (Thakkar and Chaudhari 2021) 14
2.4 London South East Services 15
2.5 Tweets containing cashtags 16
2.6 Examples of Regulatory News Statements 17

3.1 FMMS proposed by (Diaz et al. 2011) 25
3.2 Conceptual Overview of an FMMS proposed by Campbell and Keating (2013) 27
3.3 Architectural overview of an FMMS proposed by P. S. Lee, Owda, and Crockett (2018) 28

4.1 Smart Data Ecosystem Diagram 33

5.1 Joint Directors of Laboratories (JDL) Model 39
5.2 Dasarathy's fusion model (Dasarathy 1997) 41
5.3 Stock price prediction framework developed by Weng, M. A. Ahmed, and Megahed (2017) 44
5.4 Stock price prediction framework developed by X. Zhang et al. (2018) . 45
5.5 Data layer of the SDE 47
5.6 An example London South East post (not logged in) 48
5.7 An example London South East post (logged in) 48
5.8 SDE Data fusion model 50

6.1 Cashtag Collision Example 55
6.2 Cashtag collision experiment overview 59
6.3 Custom corpus creation methodology 63
6.4 Cashtag collision resolution methodology 65

7.1 Credibility Assessment Methodology (CAM) 80
7.2 Credibility feature groups 82
7.3 Top four informative features based on macro AUC 93
7.4 Sequential forward feature selection results (combined feature set) . . 96

8.1 SDE Detection Layer 100
8.2 Event generation methodology 102
8.3 Clustering of events for Boohoo PLC (LON:BOO) 105
8.4 Data points view of Boohoo PLC (LON:BOO) events 106
8.5 Clustering of tweets for Tesco (Event: Buy rating by JP Morgan 19/06/2019)107

8.6 Tweets data points view (LON:TSCO) 19/06/2019 108
8.7 Clustering of FDB Posts for Boohoo PLC (LON:BOO) for Event: Sell rating by Shore Capital 13/07/2020 109
8.8 FDB posts data points view (Event: Sell rating by Shore Capital) 110

9.1 Evaluation methodology (interview stage) 115
9.2 Participant Overview 117
9.3 Scenario 1 LSE Tweets 118
9.4 Scenario 1 Non-LSE Tweets 118
9.5 Scenario 2 LSE Tweets 119
9.6 Scenario 2 Non-LSE Tweets 119
9.7 Scenario 2 Ambiguous Tweet Text 120
9.8 Scenario 2 Ambiguous Tweet User Profile 120
9.9 Scenario 3 LSE Tweets 122
9.10 Scenario 4 - Clustering of AstraZeneca (LON:AZN) Events 124
9.11 Scenario 4 - AstraZeneca (LON:AZN) Clustering Data Points 125
9.12 Scenario 5 - Clustering of Tweets for Shore Capital Buy Broker Rating - AstraZeneca (LON:AZN) Event 14/7/2020 127
9.13 Scenario 5 - Data points View 128
9.14 Scenario 6 - Clustering of FDB Posts for Berenberg Buy Broker Rating - GlaxoSmithKline PLC (LON:GSK) Event 28/09/2020 129
9.15 Scenario 6 - Data points View 130

List of Tables

2.1 LSE Broker ratings . . . 17

5.1 SDE Data Sources . . . 47
5.2 London South East FDB post attributes . . . 49

6.1 The disparity of ticker symbols - Pets at Home Group PLC (LSE:PETS) and PetMed Express Inc (NASDAQ:PETS) . . . 55
6.2 Cashtag collision experiment data sources . . . 61
6.3 Annotated tweet examples . . . 62
6.4 Data pre-processing techniques carried out on custom corpora data sources . . . 64
6.5 Cashtag collision classifier results . . . 70
6.6 Naive Bayes results . . . 70
6.7 Logistic Regression results . . . 71
6.8 kNN Results . . . 71
6.9 DT Results . . . 72
6.10 RF Results . . . 72
6.11 SVM Results . . . 72

7.1 Financial keyword groups as defined by (Loughran and McDonald 2011) . . . 86
7.2 Inter-Item Correlation Matrix & CA Scores for binary-labelled tweets. CA = 0.591 (Sample size = 10) . . . 91
7.3 Inter-Item Correlation Matrix & CA Scores for five-class labelled tweets. CA = 0.699 (Sample size = 10) . . . 91
7.4 Inter-Item Correlation Matrix & CA Scores for three-class labelled tweets. CA = 0.686 (Sample size = 30) . . . 92
7.5 Annotated tweet breakdown . . . 92
7.6 Features removed via feature selection techniques . . . 94
7.7 Credibility classifier results . . . 96

B.1 SDE companies (Alternative Investment Market) . . . 151
B.2 SDE companies (Main Market) . . . 153

E.1 Alternative Investment Market companies (with known collisions) . . . 194
E.2 Alternative Investment Market companies (no known collisions) . . . 195
E.3 Main Market companies (no known collisions) . . . 196

F.1 General feature list . . . 197
F.2 Financial feature list . . . 199

G.1 Event clustering features . . . 202
G.2 Financial Discussion Board (FDB) post clustering features . . . 203
G.3 Tweet clustering features . . . 203

List of Abbreviations

AI	**A**rtificial **I**ntelligence
AIM	**A**lternative **I**nvestment **M**arket
API	**A**pplication **P**rogramming **I**nterface
AUC	**A**rea **U**nder **C**urve
BR	**B**roker **R**ating
CA	**C**ronbach **A**lpha
CC	**C**ashtag **C**ollision
CCRM	**C**ashtag **C**ollision **R**esolution **M**ethodology
CM	**C**onfusion **M**atrix
CSV	**C**omma **S**eperated **V**alues
CV	**C**ount **V**ectorizer
DFM	**D**ata **F**usion **M**odel
DT	**D**ecision **T**ree
EMH	**E**fficient **M**arket **H**ypothesis
EU	**E**uropean **U**nion
FCA	**F**inancial **C**onduct **A**uthority
FDB	**F**inancial **D**iscussion **B**oard
FDD	**F**inancial **D**iary **D**ate
FMMS	**F**inancial **M**arket **M**onitoring **S**ystem
FSA	**F**inancial **S**ervices **A**uthority
FTSE	**F**inancial **T**imes **S**tock **E**xchange
GDP	**G**ross **D**omestic **P**roduct
ICB	**I**ndustry **C**lassification **B**enchmark
IDS	**I**nternet **D**iscussion **S**ite
II	**I**nsider **I**nformation
IPO	**I**nitial **P**ublic **O**ffering
IT	**I**nsider **T**rading
JDL	**J**oint **D**irectors (of) **L**aboratories
JSON	**J**ava **S**cript **O**bject **N**otation
KMC	**K** **M**eans **C**lustering
KNN	**K** **N**earest **N**eighbors
LR	**L**ogistic **R**egression
LSE	**L**ondon **S**tock **E**xchange
MAD	**M**arket **A**buse **D**irective
MAR	**M**arket **A**buse **R**egulation
MC	**M**arket **C**apitilisation
MCC	**M**atthews **C**orrelation **C**oefficient
ML	**M**achine **L**earning
MM	**M**ain **M**arket
NB	**N**aive **B**ayes
NLP	**N**atural **L**anguage **P**rocessing
NLTK	**N**atural **L**anguage **T**ool **K**it
NoSQL	**N**ot only **S**tructured **Q**uery **L**anguage

NYSE	**N**ew **Y**ork **S**tock **E**xchange
OHLC	**O**pen **H**igh **L**ow **C**lose
RDBMS	**R**elational **D**atabase **M**anagement **S**ystem
RFE	**R**ecursive **F**eature **E**limination
RF	**R**andom **F**orest
RNS	**R**egulatory **N**ews **S**ervice
ROC	**R**eceiver **O**perating **C**haracteristic
SBFS	**S**equential **B**ackward **F**eature **S**election
SEC	**S**ecurities **E**xchange **C**ommission
SFFS	**S**equential **F**orward **F**eature **S**election
SDE	**S**mart **D**ata **E**cosystem
SONAR	**S**ecurities **O**bservation **N**ews **A**nalysis **R**egulation
SVM	**S**upport **V**ector **M**achine
TIDM	**T**radable **I**nstrument **D**isplay **M**nemonic
TF-IDF	**T**erm **F**requency **I**nverse **D**ocument **F**requency
UK	**U**nited **K**ingdom
UKLA	**U**nited **K**ingdom **L**isting **A**uthority

Chapter 1

Introduction

1.1 Overview

Information is gold. The ability for investors to execute well-informed investments is dependent on timely and credible information being readily available. Communication platforms such as Twitter and financial discussion boards play an essential role in enabling investors to share and assimilate stock market information. The monitoring of such information exchange has become an increasingly important aspect of ensuring the fairness and transparency of stock markets (Zaki, Theodoulidis, and Diaz 2019). This thesis presents the work undertaken relating to the research, design, and development of a Smart Data Ecosystem (SDE) to monitor stock discussion for irregular behaviour. The challenges associated with monitoring stock discussion is the vast quantity of supplementary information that investors use, such as: broker analyst ratings, financial news articles, and company reports. The research in this thesis attempts to tackle these challenges by creating an ecosystem that takes in data from multiple data sources, and combines them to enrich the analysis of discussion relating to stocks.

1.2 Motivation

Although information sources such as financial news articles and broker analyst ratings are available from trusted sources (e.g. Financial Times), information circulating on social media is less regulated and should be subjected to more scrutiny. The lack of scrutiny applied to information propagating on social media is likely to make inexperienced investors susceptible to apocryphal information and make investments based on such misinformation. Individuals responsible for spreading false information about stocks often operate with impunity, as the existing resources for monitoring such discussion are scarce. While many users who communicate on discussion channels and forums do so with the intent to inform, discuss, and assist other investors, some users are motivated by the desire for personal gain - seeking to manipulate the flow of information for short-term personal gains (Campbell and Keating 2013). The motivation behind this research is to provide mechanisms to monitor such discussion to ensure that financial markets continue to be a fair playing field for all participants.

1.2.1 Efficient Market Hypothesis

Information drives the stock market. This statement is referred to as the *Efficient Market Hypothesis* (EMH). Fama (1970) was amongst the first financial theorists to provide a definition of the EMH. The EMH is the belief that the price of a stock will always fully reflect all available information on that stock. The EMH has been scrutinised and critically reviewed for decades due to principles that undermine its theory, particularly market irregularities and theoretical paradoxes (Leković 2018).

1.3 Research Aim

This project aims to investigate and develop a novel multi-layered open-source Smart Data Ecosystem (SDE) to detect irregular behaviour relating to stocks listed on the London Stock Exchange. In the context of this research, irregular behaviour on the part of investors could include investors using specific terminology that is deemed to indicate suspicious trading activity, or perhaps posting messages in an unusual way (e.g. volume of messages or timing of messages). A smart data ecosystem can be defined as a collection of inter-working tools and platforms that have some degree of symbiotic relationship (Manikas and Hansen 2013). The layers in this ecosystem will cooperate, anonymously and automatically, to collect, clean, and analyse data from multiple communication channels to identify potentially irregular comments pertaining to financial stocks. The SDE will utilise data fusion techniques, such as time-slice windows, to provide synchronicity for the different data sources.

1.4 Research Objectives

The objectives of the research presented in this thesis are as follows:

1. Conduct research and review the variety of online communication channels used when discussing financial stocks. Research, review and compare the literature on state-of-the-art methods for detecting irregularities in different communicative environments.
2. Investigate methods to enable the creation of a data fusion model, such as time-slice windows.
3. Define the data layer of the smart data ecosystem from the data fusion model by developing the strategies to dynamically collect data from various channels. Populate the smart data ecosystem with twelve months of data.
4. Develop and deploy classifiers to assist the data management process in respect to assessing the credibility of microblogging posts and resolving naming conflicts present in data sources
5. Design and implement the detection layer of the smart data ecosystem by developing and deploying unsupervised clustering algorithms to identify potentially irregular events and posting activity.
6. Evaluate the level of assistance of the ecosystem through a set of example scenarios.

1.5 Research Questions

Four research questions are addressed in this thesis, each of which are addressed in various chapters:

1. Can a smart data ecosystem, utilising machine learning classifiers, accurately classify social media posts with respect to their credibility? (Chapter 7)
2. Can a smart data ecosystem be used to automate the monitoring of a variety of communication channels for irregular behaviour? (Chapter 8)

3. Can a smart data ecosystem, utilising clustering algorithms, identify irregular days and events with respect to posting activity? (Chapter 8)
4. Can a smart data ecosystem, through visualisation tools, assist a user in establishing the significance of detected irregularities? (Chapter 9)

1.6 Contributions

The main contributions of this thesis can be summarised as follows:

- A novel **data fusion model** (Chapter 5) for the fusing of heterogeneous financial market data sources. This data fusion model addresses the challenges of combining multiple data sources into one unified unit, dealing with timestamp refinements, and resolving cashtag collisions.
- A **novel methodology for resolving cashtag collisions on Twitter** (Chapter 6). This methodology combines data from different sources to develop company-specific corpora, which motivates various features to train machine learning models to classify tweets as related to a specific exchange or not.
  - Machine learning models capable of classifying a tweet as being related to a specific exchange or not - an issue not addressed until this research was undertaken. These classifiers were trained on a manually-annotated dataset of 5,000 tweets. Accuracy scores of up to 94.9% were obtained in experiments to validate the methodology.
- A **novel methodology for assessing the credibility of financial stock tweets** Chapter 7). This methodology involves the training of machine learning models to classify the credibility of tweets.
  - Machine learning models capable of classifying a financial stock tweet as being one of three classes: (1) not credible, (2) ambiguous, or (3) credible. These classifiers were trained on a manually annotated data of 5,000 tweets, using general features found in all tweets, and novel financial features typically found in tweets relating to stocks. Classifiers achieved scores of up to 94.3 AUC.

- A **smart data ecosystem** which monitors for irregular discussion and events. This ecosystem utilises all of the previous contributions listed. The effectiveness of the aforementioned contributions are evaluated by conducting a series of qualitative interviews with financial markets experts (Chapter 9).

1.7 Thesis Overview

The research presented in this thesis is presented over ten chapters. Figure 1.1 provides a high-level overview of how the objectives motivate certain chapters, and the research outputs which serve as the motivation for their respective chapters.

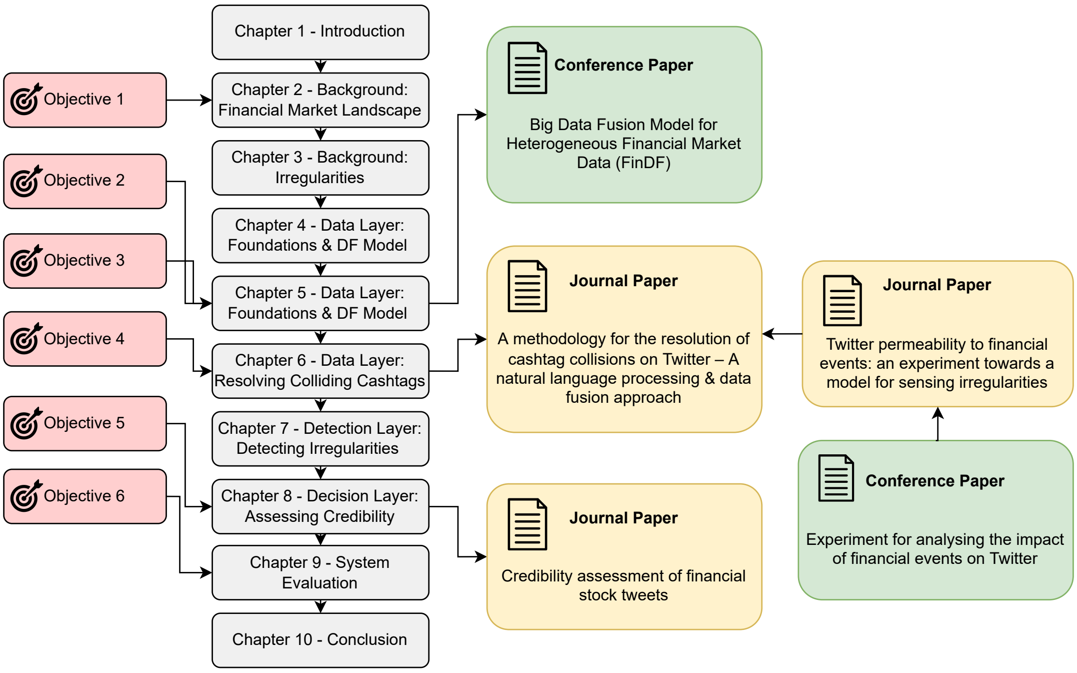

Chapter 1 - Introduction
Chapter 2 - Background: Financial Market Landscape
Chapter 3 - Background: Irregularities
Chapter 4 - Data Layer: Foundations & DF Model
Chapter 5 - Data Layer: Foundations & DF Model
Chapter 6 - Data Layer: Resolving Colliding Cashtags
Chapter 7 - Detection Layer: Detecting Irregularities
Chapter 8 - Decision Layer: Assessing Credibility
Chapter 9 - System Evaluation
Chapter 10 - Conclusion
Objective 1
Objective 2
Objective 3
Objective 4
Objective 5
Objective 6
Conference Paper
Big Data Fusion Model for Heterogeneous Financial Market Data (FinDF)
Journal Paper
A methodology for the resolution of cashtag collisions on Twitter – A natural language processing & data fusion approach
Journal Paper
Twitter permeability to financial events: an experiment towards a model for sensing irregularities
Conference Paper
Experiment for analysing the impact of financial events on Twitter
Journal Paper
Credibility assessment of financial stock tweets

FIGURE 1.1: Links between the research objectives, thesis chapters, and research outputs

Chapter 2 provides a background on financial markets, including a breakdown of financial market participants and the data sources investors rely on to make decisions.

Chapter 3 introduces the notion of irregularities within financial markets and the techniques employed to spot such irregularities.

Chapter 4 details the high-level research methodology. The Smart Data Ecosystem (SDE) is formally introduced in this chapter, including the role each of the layers play in the SDE.

Chapter 5 introduces the data layer of the SDE. This chapter begins with outlining the related work on the fusing of stock market data. The chapter then introduces a conceptual model for fusing heterogeneous financial market data within the ecosystem.

Chapter 6 presents a methodology that aids the data fusion model to detect tweets that contain colliding cashtags. The outcome of this methodology is a classifier capable of classifying a tweet as belonging to a specific exchange or not; in the case of this thesis, the London Stock Exchange.

Chapter 7 introduces the methodology for assessing the credibility of financial stock tweets. This chapter will firstly introduce the related work on assessing the credibility of tweets, before introducing the ecosystem's classifiers for assessing credibility, including the features used.

Chapter 8 introduces the detection layer of the SDE, which utilises unsupervised clustering algorithms to identify irregular events and the discussion taking place within events.

Chapter 9 discusses the evaluative study undertaken of the various ecosystem tools developed during this research. This evaluation involved interviewing five financial market experts to ascertain the effectiveness of ecosystem to detect irregularities.

Chapter 10 concludes with a summary of the contributions of the work undertaken and proposes avenues for future work.

Chapter 2

Background: Financial Market Landscape

2.1 Overview

This chapter provides an overview of financial markets, with particular emphasis on the UK's stock market - the London Stock Exchange. The purpose of this chapter is to set the scene and provide context into the key participants of the financial marker, and how they are involved in this research.

Not every stock market is alike - some are formed of different sub-markets which have different listing requirements and regulations. Companies can also be listed on an index - a list of companies bundled by their market capitalisation. This chapter begins by providing a brief history of financial markets since their inception (Section 2.2). The stock market that is the focus of this research, the London Stock Exchange, is then introduced in Section 2.3. The key participants involved within the financial market are then introduced in Section 2.4. Next, several of the most prominent data sources that investors use when performing analysis and discussion of stocks are discussed in Section 2.5.

2.2 A Brief History of Financial Markets

Stocks markets allow companies to raise long-term capital through the selling of shares, which represents partial ownership of a company. Stock markets can trace their history back over four hundred years. The Amsterdam Stock Exchange (now known as the Euronext Amsterdam) is considered to be the first official stock exchange, with trading commencing in 1602. There are now sixty major stock exchanges across the world (Istiake Sunny, Maswood, and Alharbi 2020). The 1990s were a remarkable decade for stocks; the Dow Jones and S&P indices (the top 30 and 500 US-listed companies respectively in terms of market capitalisation) rose by over 400% (Mishkin 2016).

2.3 London Stock Exchange

The London Stock Exchange (LSE) is the stock exchange of the United Kingdom (UK), and is both a primary market and a secondary market. Organisations can raise capital by selling shares to investors on the primary market, and shares can then be continuously traded between investors on the secondary market. Although the thought of a stock market often invokes images of traders frantically running across a trading floor, the LSE was amongst one of the first stock exchanges to abandon its traditional floor trading in the 1980s and move towards complete digitisation (Clemons and Adams 1988).

One of the most monumental actions a company can take is to list itself on the stock market - often referred to as ″going public″ - this is officially known as the company’s Initial Public Offering (IPO). The primary purpose of an IPO is for a company to raise funds for future growth. Companies listed on the LSE are either listed on the Main Market or the Alternative Investment Market after going public, each of these markets have different requirements and purposes, which will now be discussed.

2.3.1 Main Market

The Main Market (MM) of the LSE, also referred to as the *official list*, is the primary sub-market of the LSE. Over a thousand companies are listed on the MM at any given time. Companies wishing to join the MM have to follow a two-step process; their securities (shares) have to be:

- admitted to the official list by the United Kingdom Listing Authority (UKLA) - part of the Financial Conduct Authority (the LSE regulator)

- admitted to the official list by the exchange for trading

Companies obtaining a position on the MM must ensure that at least twenty five per cent of their share capital is in public hands, so that the shares are capable of being actively traded on the market (Arnold 2014). Companies that do not meet the requirements of listing on the MM may choose to list themselves on the smaller sub-market of the LSE - the Alternative Investment Market.

2.3.2 Alternative Investment Market

Companies which do not meet the stringent requirements of being listed on the MM may be able to list on LSE's Alternative Investment Market (AIM). The AIM was opened in 1995, as a result of a long-recognised need for small companies to raise equity capital (ibid.).

Although the AIM is predominately known for housing smaller, less well-known companies, that is not strictly always the case. For example, in 2016 the clothing company ASOS PLC boasted a market capitalisation of £5bn - much higher than many of its MM counterparts (Doukas and Hoque 2016). In recent times, more firms which meet the heavier regulatory requirements of the MM have decided to list themselves on the lighter regulatory environment of the AIM (ibid.). Doukas and Hoque (ibid.) found that almost half of the companies that issued equity on the AIM met the requirements to list on the MM, but chose not to. The research undertaken by Doukas and Hoque (ibid.) posited the key reasons for this is that smaller, less-established, companies choose to continue issuing equity on AIM due to the lower listing and ongoing costs and that companies choose their market based on their financing and growth strategy.

2.4 Stock Market Participants

Although there are many participants that allow the seamless operation of stock markers; there are six key participants that are central to understand for the purposes of this research. These key participants (Figure 2.1) are; companies, investors, brokers, government, the stock market itself, and regulators within the country the

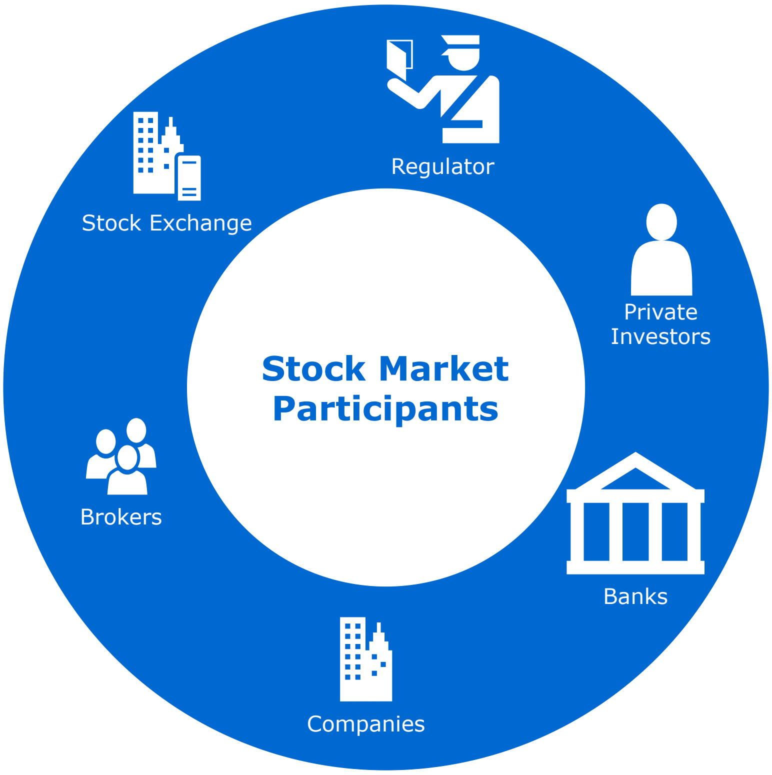


FIGURE 2.1: Stock Market Participants

stock market operates. This section will detail each of these key participants, with specific emphasis on how they play a part in how the LSE operates.

2.4.1 Companies

Companies are one of the primary participants of the stock market. Shares made available by the company (through an IPO) are initially traded on the primary market, which can then be traded between investors on the secondary market.

When undertaking an analysis of company performance, a common task is to compare a company to others within the same industry or sector. Companies listed on the LSE belong to a specific sector and industry as outlined in the Industry Classification Benchmark (ICB). The ICB is a globally recognised standard for categorising companies by industry and sector. The ICB operates a four-tier structure that encompasses 11 industries, 20 super-sectors, 45 sectors, and 173 sub-sectors. The ICB has been adopted by other exchanges worldwide, including the NYSE, NASDAQ, and Euronext (FTSE Russell 2021).

2.4.2 Investors

Investors can take many forms - private individuals, banks, governments, and even other companies. Investors are typically categorised into two types based on their investing behaviour - passive or active. As the name implies, passive investors raise their capital passively over time, and are not likely to take risks. Active investors are the opposite - they actively trade on information and have a high risk tolerance. Active investors are found to be more experienced due to their more active investment strategy (Jureviciene and Jermakova 2012).

Although it is perfectly acceptable for investors to deal directly with one another off the exchange, the majority of trades occur through brokers who act on behalf of investors (Arnold 2014).

2.4.3 Broker Agencies

There are currently over a hundred brokers which provide analysis of LSE-listed companies (London Stock Exchange 2021). These broker agencies provide public ratings - opinions that fall broadly into buy, hold, and sell categories - to reflect the broker′s opinion on a specific stock, and also carry out investments on behalf of their clients.

2.4.4 Banks

There are four general aspects of the banking system. Firstly, high street banking refers to services provided to the general public. Secondly, business banking relates to specialised services afforded to businesses. Thirdly, central banks, typically a quasi-government establishment, ensure that there is sufficient liquidity in the market. The Bank of England, the UK's central bank, is responsible for monitoring and adjusting interest rates, ensuring stable economic growth, and aims to keep inflation low. Finally, investment banking refers to financial institutions that invest money on high street banks, investment trusts, and pension funds.

Investment banks will attempt to invest money through knowledge of the stock market and assist companies involved in mergers and acquisitions. All activity within investment banks takes place on either the ″sell-side″ or the ″buy-side″. The "sell side" is concerned with helping companies raise debt, which will ultimately be sold to investors such as hedge funds and insurance companies. The "buy-side" relates to institutions that buy shares for money-management purposes.

2.4.5 Regulator: Financial Conduct Authority

Following the 2007/08 financial crisis, the UK government decided that existing regulation of financial markets were not adequate. The Financial Services Authority (FSA) were originally responsible for the regulation of financial markets. In 2012, the Financial Services Act Financial Services Act 2012 (2012) was introduced in an attempt to give more authority and control to the regulators of financial markets. The act led to the creation of the Financial Conduct Authority (FCA), which superseded the FSA. Research by Pham and Ausloos (2020) has found that, following the introduction of the Financial Services Act 2012, prices are less noisier, and that the FCA is efficient in regulating insider trading.

As the current market regulator of the LSE, the FCA (Financial Conduct Authority 2019) enforcement powers include:

- withdrawing a firm's authorisation to trade
- prohibiting individuals from carrying on regulated activities
- suspending firms and individuals from undertaking regulated activities
- issuing fines against firms and individuals who breach the FCA's rules or commit market abuse
- issuing fines against firms breaching competition laws
- making a public announcement when the FCA begin disciplinary action and publishing details of warning, decision and final notices
- applying to the courts for injunctions, restitution orders, winding-up and other insolvency orders
- bringing criminal prosecutions to tackle financial crime, such as insider dealing, unauthorised business and false claims to be FCA authorised

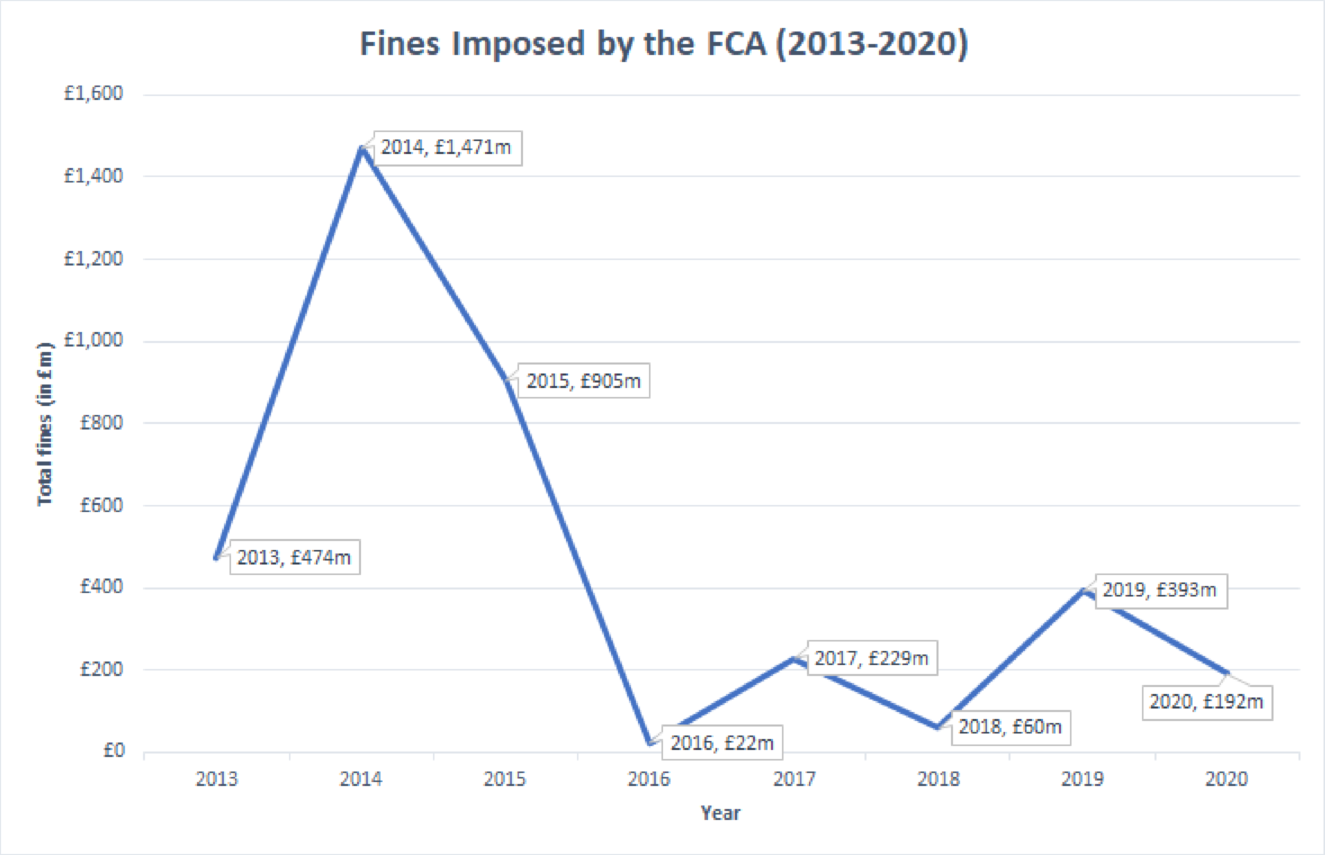


FIGURE 2.2: Fines issued by the FCA (2013 - 2020) (Financial Conduct Authority 2021)

- issuing warnings and alerts about unauthorised firms and individuals and requesting that web hosts deactivate associated websites

Since the inception of the FCA in 2012, over two hundred fines have been imposed up to the end of 2020 (Figure 2.2) (Financial Conduct Authority 2021). Breaking these fines down, 118 of them have been imposed against companies, with the remaining 93 imposed against individuals, highlighting the regulator's interest in pursuing both companies and individuals for market misconduct.

2.5 Stock Market Information Sources

The advent of the internet means investors are no longer dependent on information sources such as the morning newspaper for information about stocks. Instead, financial discussion boards, social media, and broker analyst ratings available online are just a sample of sources investors can digest for information.

Analysis of stock information generally falls under fundamental analysis or technical analysis (Suresh 2013). Fundamental analysis is concerned with attempting to forecast future price movements in an attempt to profit from such movements. Fundamental analysis is not limited to analysis of stocks, but also considers the overall economy and industry conditions. Technical analysis is usually used to supplement fundamental analysis, as opposed to a substitute to it. This type of analysis focuses on statistical trends relating to a stock's volume and price. Charts are typically used in this type of analysis to identify trends which suggest how a stock will perform in the future. Figure 2.3 provides an overview of the data sources used for fundamental analysis, technical analysis, crowd-sourced data sources, and quantitative data used within the investment community. This section will detail some of the most prominent information sources used by investors to facilitate discussion.

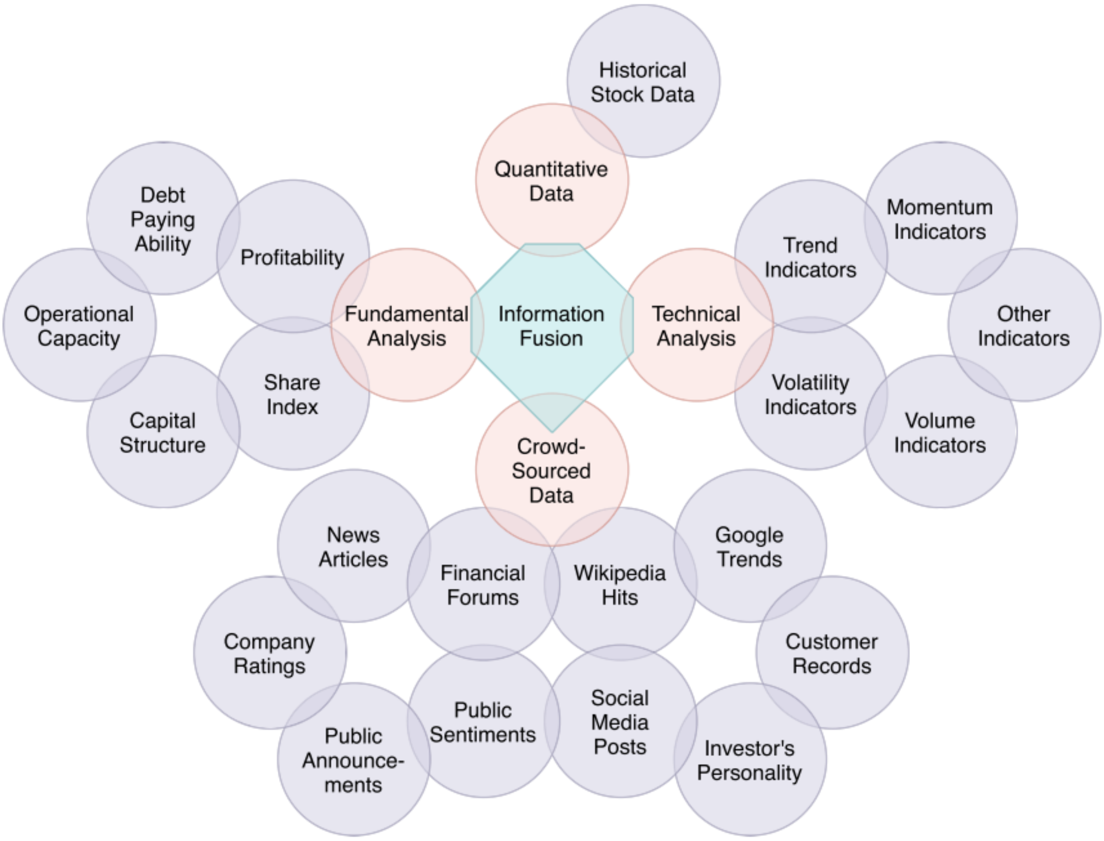
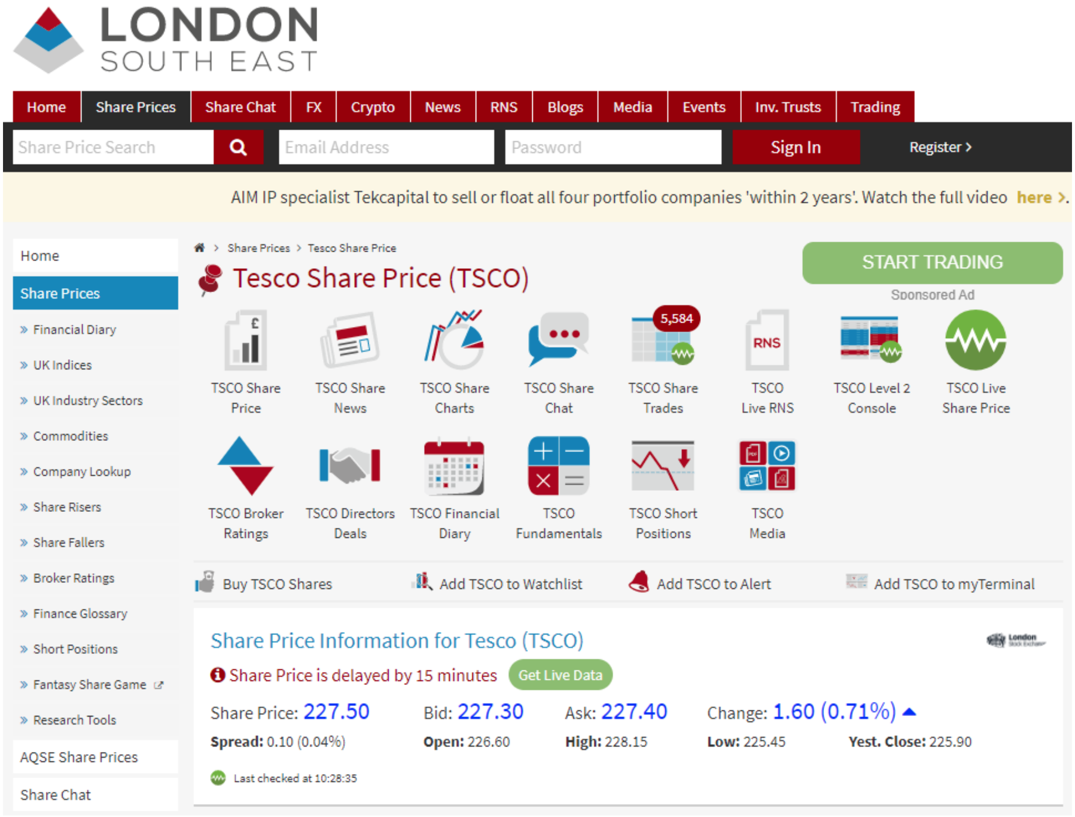


FIGURE 2.3: Financial data sources (Thakkar and Chaudhari 2021)

2.5.1 Financial Discussion Boards (e.g. London South East)

Existing before the emergence of social media platforms such as Twitter, financial discussion boards (FDBs) have provided a place in which investors can disseminate information relating to stocks. Many FDBs provide a dedicated sub-forum for each stock-listed company. Many FDBs are self-regulated, in which administrators of the website will monitor discussion, assisted by users who report inappropriate or misleading content to be reviewed (Campbell and Keating 2013).

Prominent FDBs which focus on LSE-listed stocks include London South East[1], Interactive Investors[2], and ADVFN [3]. Many of these FDB also aggregate many other types of information (Figure 2.4), such as providing historical company accounting data, broker ratings, and important dates for companies (e.g. dividend dates).



AIM IP specialist Tekcapital to sell or float all four portfolio companies 'within 2 years'. Watch the full video here >.

Tesco Share Price (TSCO)

Share Price Information for Tesco (TSCO)

Share Price is delayed by 15 minutes

Share Price: 227.50 Bid: 227.30 Ask: 227.40 Change: 1.60 (0.71%) ▲

Spread: 0.10 (0.04%) **Open:** 226.60 **High:** 228.15 **Low:** 225.45 **Yest. Close:** 225.90

Last checked at 10:28:35

FIGURE 2.4: London South East Services

[1] https://www.lse.co.uk/

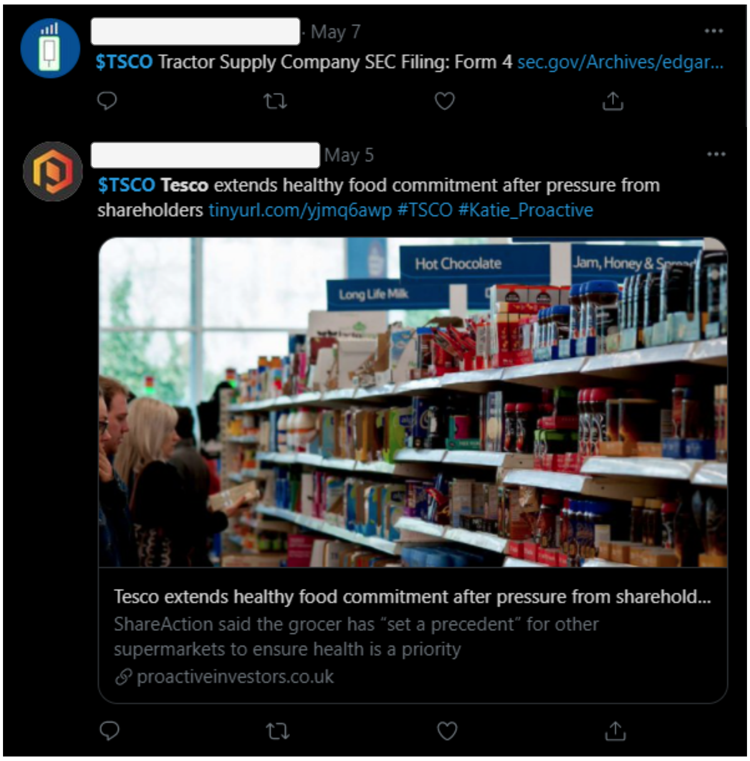
[2] https://www.ii.co.uk/

[3] https://uk.advfn.com/

2.5.2 Social Media (e.g. Twitter)

Allowing for the fast dissemination of information, micro-blogging websites such as Twitter have become increasingly used by investors to gather news relating to stocks. Recognising the increasing demand of stock discussion in 2012, Twitter introduced a feature for users to align their tweets with specific companies - the cashtag. Similar in design to a hashtag (prefixing a word with a # symbol to create a clickable tag), a cashtag can be created by prefixing a company's ticker symbol with a $ symbol (Cresci, Fabrizio Lillo, et al. 2018). Words used to form cashtags mimic a company's Tradable Instrument Display Mnemonics (TIDM) - a series of characters unique to that company on the exchange they are listed on.

Cashtags suffer from a key drawback, in that companies listed on different exchanges may possess an identical TIDM to a company listed on another exchange. Figure 2.5 illustrates this phenomenon - the TIDM for Tesco PLC on the LSE is TSCO, likewise, on the NASDAQ, the Tractor Supply Company is also TSCO. When investors search for such cashtags, Twitter does not differentiate between them - potentially sowing confusion for investors who use such information sources.



[redacted] · May 7 ···

$TSCO Tractor Supply Company SEC Filing: Form 4 sec.gov/Archives/edgar...

[redacted] May 5 ···

$TSCO **Tesco** extends healthy food commitment after pressure from shareholders tinyurl.com/yjmq6awp #TSCO #Katie_Proactive

Tesco extends healthy food commitment after pressure from sharehold...

ShareAction said the grocer has "set a precedent" for other supermarkets to ensure health is a priority

proactiveinvestors.co.uk

FIGURE 2.5: Tweets containing cashtags

2.5.3 Broker Ratings (via Broker Agencies)

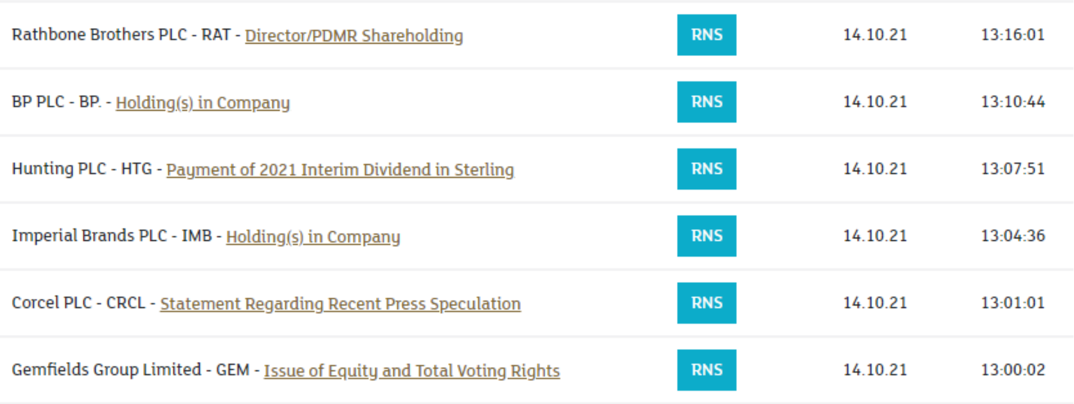
Broker agencies provide broker ratings after undertaking analysis of a particular stock-listed company. These ratings generally fall into buy, hold, or sell groups and aim to predict an equity's future performance (Premti, Garcia-Feijoo, and Madura 2017). Table 2.1 shows the different types of ratings which can be issued by brokers on the LSE - although some brokers may only use a subset of these.

TABLE 2.1: LSE Broker ratings

Buy (Positive)	**Hold (Neutral)**	**Sell (Negative)**
Strong Buy	Hold	Sell
Buy	Maintain	Reduce
Accumulate	Neutral	Unattractive
Overweight	Market Perform	Underweight
Outperform	In-line	Underperform
		Strong Sell

2.5.4 Regulatory News Services (RNS)

The Regulatory News Service (RNS) of the LSE provides a platform in which important company announcements and price-sensitive news is made available to investors by stock-listed companies(Arnold 2014). Each RNS announcement is associated with a company listed on the LSE, and also contains a title that summarises what the RNS relates to. Examples of RNS announcements (Figure 2.6) include company leadership changes (e.g. new CEO appointment), addressing speculation around rumours that may be circulating, and results of annual and emergency general meetings.



Rathbone Brothers PLC - RAT - Director/PDMR Shareholding	RNS	14.10.21	13:16:01
BP PLC - BP. - Holding(s) in Company	RNS	14.10.21	13:10:44
Hunting PLC - HTG - Payment of 2021 Interim Dividend in Sterling	RNS	14.10.21	13:07:51
Imperial Brands PLC - IMB - Holding(s) in Company	RNS	14.10.21	13:04:36
Corcel PLC - CRCL - Statement Regarding Recent Press Speculation	RNS	14.10.21	13:01:01
Gemfields Group Limited - GEM - Issue of Equity and Total Voting Rights	RNS	14.10.21	13:00:02

FIGURE 2.6: Examples of Regulatory News Statements

2.6 Chapter Summary

This chapter has presented an overview of how financial markets operate, focusing primarily on the LSE. The key participants that ensure the smooth operation of financial markets were also introduced, including the the role of each of these participants. It is important to understand that many factors and participants are at work, each of which have an impact on stock prices and the discussion surrounding them. Natural disasters, pandemics, and financial crashes are just a number of examples that can rock the fragile financial market landscape, and could potentially give rise to participants such as private investors to attempt to manipulate share prices for their personal gain.

With the background to financial markets discussed, the next section will detail the related work relating to this research: the detection of financial market irregularities.

Chapter 3

Background: Detection of Irregularities

3.1 Overview

This chapter will provide an overview of irregularity detection in the context of detecting financial market irregularities. The primary purpose of this chapter is to determine which common characteristics exist between the different existing frameworks when it comes to detecting irregular behaviour within the financial market. This chapter will provide the necessary background information on irregularity detection focusing specifically on financial markets, before the research methodology is discussed in Section 4. Firstly, a general definition of an irregularity is provided (Section 3.2). Several well-documented financial market irregularities are then introduced (Section 3.3). The background on irregularity detection is then discussed (Section 3.4), followed by the techniques adopted within the literature to detect financial market irregularities (Section 3.5).

3.2 Defining an Irregularity

An irregularity is synonymous with an anomaly, with anomalies defined as "patterns in data that do not conform to a well-defined notion of normal behaviour" (Chandola, Banerjee, and V. Kumar 2009). This thesis will use the term irregularity in order to maintain consistency. Irregularities indicate significant and rare events, often demanding the attention of an expert in the given domain when they are discovered. For example, irregular credit card activity could indicate credit card fraud, and network traffic patterns that do not conform to the normal observed behaviour could signify that a computer system is under attack (M. Ahmed, Mahmood, and Islam 2016). Irregularities are not noise in the data, as noise is often meaningless and ignored or removed from datasets – irregularities, on the other hand, translate to significant (and often critical) actionable information (Chandola, Banerjee, and V. Kumar 2009).

3.3 Documented Financial Market Irregularities

The long-standing nature of financial markets means that several irregular events have been documented since their existence. Schulmerich, Leporcher, and Eu (2015) group financial market-based irregularities into four categories:

1. **Fundamentals** - is a type of irregularity that is noticed through the study of accounting data. One example being the price-to-earning (P/E) ratio effect. Research has found that low P/E stocks tend to outperform both the market and high P/E stocks (ibid.), and the study of such accounting data can be exploited.

2. **Calendar** - is a type of irregularity that refers to those scenarios where stocks appear to perform differently depending on the time of the year. The most well-documented calendar-based irregularity being the January Effect, whereby stock prices have a tendency to rise during the month of January.

3. **Structure-related** - is a type of irregularity that relates to market transparency, how a specific market is regulated, and unfair competition. A well-known irregularity of this type is the Merger Arbitrage anomaly, in which the value of the company being acquired (as part of a merger and acquisition process) tends to rise while the value of the acquiring firm tends to decline.

4. **Behaviour-based** - is a type of irregularity that includes brokers who generate trading patterns that could potentially affect the market and the behaviour of investors. The most prominent type of behaviour-based irregularity include

insider trading, in which insiders of a company (typically executive-level persons) trade as a result of knowing information not yet released to the public, giving them an unfair advantage when trading over investors who are not privvy to such information.

This research focuses primarily on behaviour-based irregularities, particularly focusing on the communicative behaviour of investors over social media and FDBs, as the emergence of social media platforms such as Twitter provides a new dimensional to how investors behave (Nofer and Hinz 2015).

The use of Artificial Intelligence (AI) and Machine Learning (ML) in the stock market sector has been heavily geared towards the prediction of stock prices (Y. Kim and Sohn 2012), rather than identifying irregularities that could be indicative of stock market manipulation (Close and Kashef 2020). Existing work of irregularity detection within the sphere of financial markets places emphasis on price movements, with few pieces of work looking at what is being *discussed*, particularly by investors on discussion sites.

In the context of this research, an irregularity could present itself within the activity of a certain stock at a certain day. A company that typically is not discussed in any great depth by investors which suddenly sees a spike in discussion across different discussion channels could be considered irregular in the context of that company. It may be perfectly typical for a company to not be very well discussed on platforms such as Twitter and FDBs, but then suddenly see a surge in posting activity at specific periods of the year.

3.4 Irregularity Detection in Financial Markets

Irregularity detection has been employed in a variety of domains to detect irregular patterns that deviate from the normal expected behaviour (Chandola, Banerjee, and V. Kumar 2009). Areas such as fault diagnosis, intrusion detection systems, and fraud detection have benefited from advancements in irregularity detection (Hayward and Madill 2004).

According to M. Ahmed, Mahmood, and Islam (2016), irregularities can be categorised into three distinct groups:

1. **Point irregularity** – a particular data instance that deviates from the normal pattern of the dataset, it can be considered as a point irregularity.
2. **Contextual irregularity** – a data instance is behaving irregular in a particular context, but not in another context. In the financial market, this is similar to calendar-effect irregularities such as the January effect.
3. **Collective irregularity** – When a collection of similar data instances is behaving irregular with respect to the entire dataset, then this collection is termed as a collective irregularity.

3.4.1 Irregularity Detection Challenges

Although irregularity detection is now a well-establish field of research with applications in many disciplines, it is not without its challenges. M. Ahmed, Mahmood, and Islam (2016) and Chandola, Banerjee, and V. Kumar (2009) state the principal challenges of irregularity detection:

- Supervised approaches of irregularity detection require labelled data - which is scarcely available
- Malicious users attempt to make irregular behaviour appear normal by imitating normal activities – often circumventing detection mechanisms
- Normal behaviour typically changes over time - what is considered typical behaviour now may be atypical in the future
- Irregularities are often specific to the context - what is irregular in one scenario (e.g. a specific company) could be typical behaviour in another
- Irregularity detection techniques can be difficult to generalise to other domains - an irregularity detection methodology for detecting intruders over a computer network may face challenges in being deployed in other areas

3.5 Financial Market Monitoring

For the purposes of this thesis, financial market monitoring follows the definition presented in Polansky, Kulczak, and Fitzpatrick (2004), in which market surveillance is defined as "the processes and technologies that support the detection and investigation of potential trading rule violations, whether defined in statute or marketplace rules". A Financial Market Monitoring System (FMMS) can therefore be understood as the sub-set of processes and technologies to support the detection and investigation of stock market fraud and irregular behaviour (Diaz et al. 2011). Several attempts to design and develop Financial Market Monitoring Systems (FMMS) to monitor financial markets have been proposed over the years. Such systems may monitor specific elements of the financial market, such as the movement of stock prices (Y. Kim and Sohn 2012), irregular comments posted on FDBs (Owda, Crockett, and P. S. Lee 2017; P. S. Lee, Owda, and Crockett 2018), or posting activity surrounding certain stocks (Sabherwal, Sarkar, and Y. Zhang 2011).

This section will firstly introduce FMMSs which have been proposed in the literature for monitoring financial market irregularities and fraud. Following the overview of conceptual frameworks, an overview of statistical and machine learning models which have supported the monitoring of financial markets are then discussed.

3.5.1 Proposed FMMS Frameworks

Several FMMS have been proposed over the years. Many of the existing systems used by regulatory bodies that are currently in existence are not well-documented within the literature, as regulators utilising such systems are wary of exposing methodological processes that could allow manipulators to circumvent such systems.

One of the earliest FMMS discussed within the literature – and ultimately deployed – to monitor stock market data sources was the Securities Observation, News Analysis, and Regulation (SONAR) system. This system was created to monitor the NASDAQ stock exchange, including Over the Counter (OTC) and NASDAQ-Liffee (futures) stock markets for potential insider trading and fraud by way of misrepresentation. Developed by the National Association of Securities Dealers (NASD) and operational since December 2001, this system utilised heterogeneous data sources to effectively monitor stock markets for potential insider trading and fraud. The data sources this system considered included news wire stories, Securities and Exchange Commission (SEC) filings, and stock prices. The SONAR system was capable of processing 10,000 news stories and SEC filings and generated 50-60 alerts per day (Goldberg et al. 2003). These alerts were ultimately reviewed by regulatory investigators for a final assessment of the severity of the alert.

According to Goldberg et al. (ibid.), the SONAR system combined components in order to:

- detect evidence as it occurs in text sources (news wires and SEC filings)
- detect characteristic ”events” in a space of price/volume-derived feature of market activity
- combine this evidence in a meaningful way by assigning a probability-like score to each ”security-day” which estimates the likelihood of several episodes of regulatory interest.

Although this system was implemented two decades ago, it successfully employed AI and statistical techniques, which included natural language processing (NLP) text mining, rule-based inference, fuzzy matching, and statistical regression. The system was evaluated against the system currently in use at that time, Stock Watch Automated Tracking (SWAT), and the time taken for a human investigator to review an alert ranged from 15-20 minutes (SONAR), versus 30-60 minutes (SWAT).

Diaz et al. (2011) presented a systematic framework (Figure 3.1) of an FMMS that considers data from a variety of sources and produces alerts based on potential irregular activity. The specific data sources considered by this system were not provided, but the authors did note that data could include: intraday share prices (e.g. open, high, low, close prices of a stock), company profile information (e.g. employee information), along with financial statements relating to stock-listed companies, and textual sources such as financial news, internet forums, blogs, and financial events in the form of filings with the relevant authorities, such as the Securities and Exchange Commission (SEC) in the United States. The components for analysis took two forms: behaviour analysis and economic analysis. Behaviour analysis contained

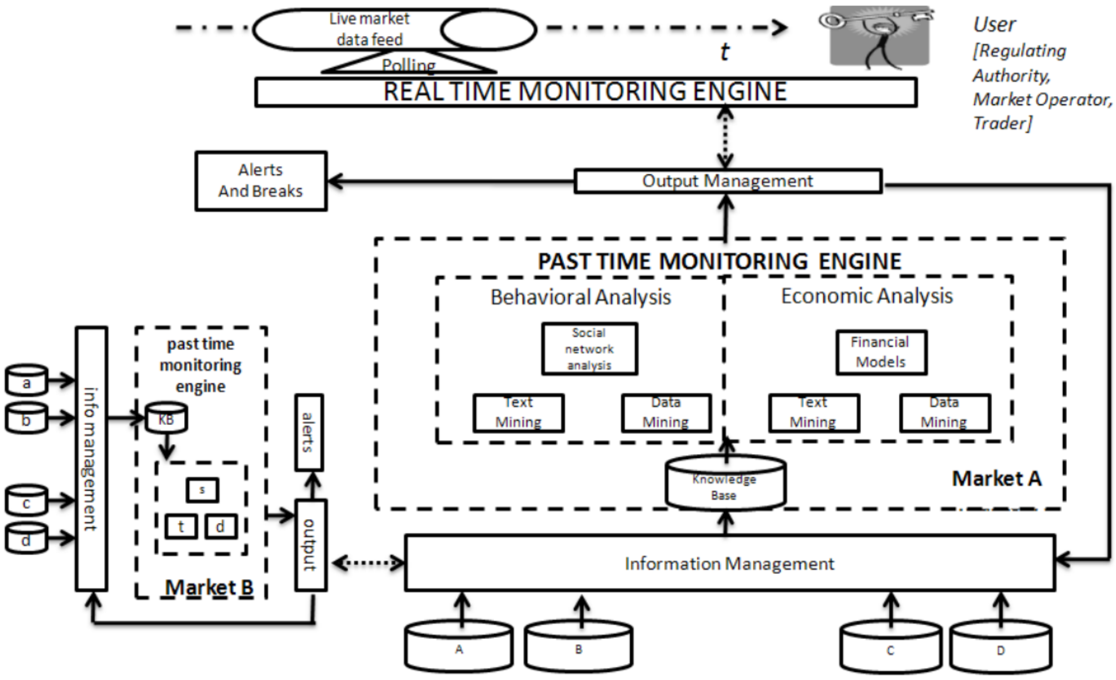


FIGURE 3.1: FMMS proposed by (Diaz et al. 2011)

sub-components for social network analysis, text mining, and data mining. The economic analysis included modules for financial modelling, data mining modelling, and text mining analysis. As the framework proposed by Diaz et al. (ibid.) was conceptual, challenges relating to implementing such a system were only partially explored and discussed. The work of Diaz et al. (ibid.) differs to the research presented in this thesis in that none of the user-generated data sources (e.g. Twitter, Financial Discussion Board posts) to be collected and analysed are explored, nor are any of the challenges (e.g. identifying the company/stock being discussed by investors) of collecting such vast quantities of data from such sources considered. One of the primary concerns of any FMMS is how data is collected, stored, and analysed to ascertain if data point(s) are irregular or warrant further investigation by an industry professional.

Campbell and Keating (2013) proposed a conceptual model (Figure 3.2) to support the development of a decision support system to aid investors and Internet Discussion Site (IDS) administrators to monitor communicative behaviour on IDSs. Although this conceptual model presented the relationships which support information sharing within the financial market, it neglected to address logistical concerns such as stock market data collection and storage.

The project consisted of four phases. The pilot phase (1) involved undertaking a review of the academic literature relating to IDS and communicative practices within the financial market. The exploratory phase (2) involved selecting an appropriate IDS to monitor - electing to focus on an IDS which must have been in operation for at least 10 years in order to provide a sufficient pool of data. The final step of this phase involves the interviewing of key stakeholders within the IDS community to determine the relevance of the process depicted in their model. The explanatory phase (3) proposed undertaking two surveys targeted towards stock market participants (e.g. investors). The first survey aimed to build up a profile of the respondent (i.e. trading experience, risk orientation etc). The second survey presented the investors with scenarios and asked their likely response if the scenario was real. The final phase (4) involved the development of the system.

The paper provided a description of the research progress towards their system, with the first phase being completed. However, no further research has been published related to the proposed system, indicating possible methodological limitations in the development process.

A prototype Financial Market Surveillance Decision Support System (FMS-DSS) was developed by Alić (2015) that focused on detecting potential pump-and-dump manipulation that also utilised voluminous and heterogeneous data streams. The pump-and-dump manipulation scheme is one of the most well-known information-based market manipulation techniques. A user first purchases shares at the typical market price, and then proceeds to spread false positive information to market participants in the hope the share price increases as a result of the increase popularity of the stock. Once the share price rises, the manipulator can then sell their shares at a profit, before the share price dips to its original level (Siering et al. 2017). Although this research claimed to provide convincing evidence for a long-term analysis of real data, the FMS-DSS developed was not evaluated on a real-world dataset under real-world conditions.

On the subject of detecting rumours and misinformation, Majumdar and Bose (2018) proposed a framework based on knowledge-based discovery in databases and detection of fraudulent financial activities, and identified several critical factors that lead to identifying financial rumours. This research included curating a list

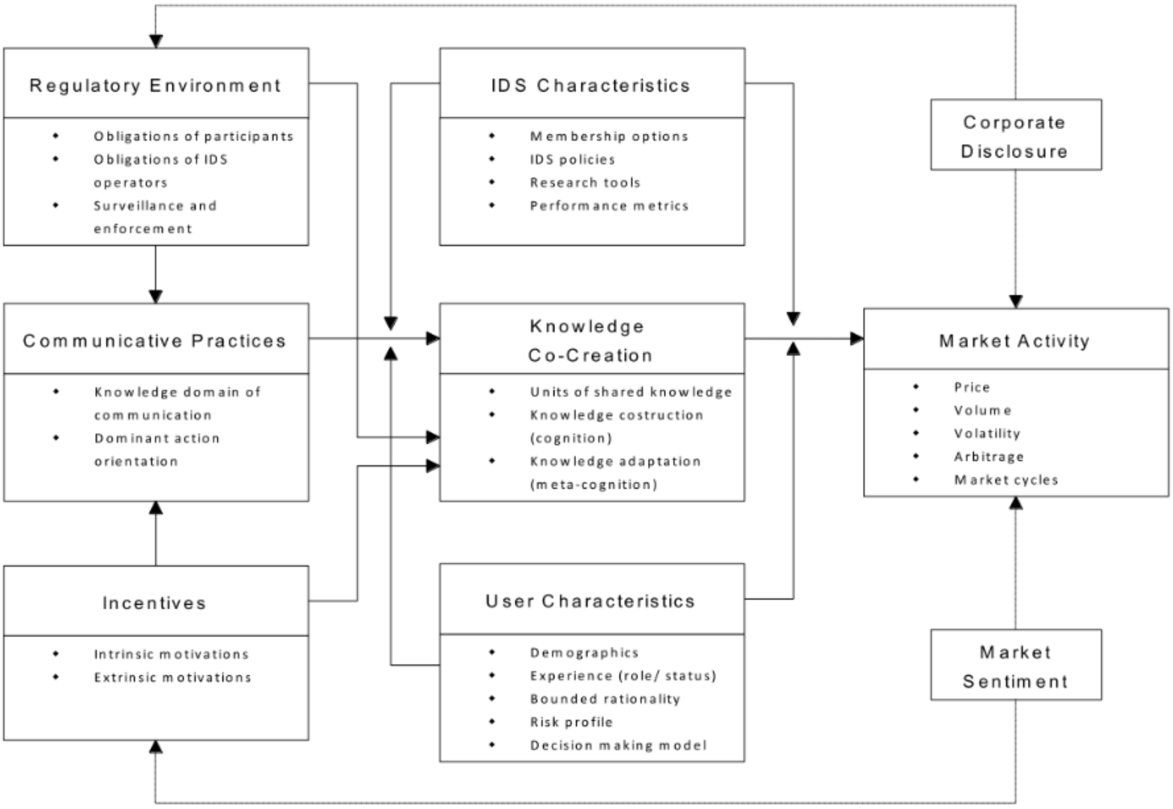


FIGURE 3.2: Conceptual Overview of an FMMS proposed by Campbell and Keating (2013)

(through consulting experts) of keywords that generally denoted a financial rumour within the data (e.g. FDB post or tweet). One of the findings relating to generating this keyword list is that the keywords of interest would vary based on the communication environment. Twitter, for example, feature a character limit of 280 characters, meaning their is an abundance of acronyms and other micro-blogging nuances such as hashtags present in such data, meaning variations of the keywords (e.g. abbreviations) needs to be adapted for the Twitter environment.

P. S. Lee, Owda, and Crockett (2018) proposed a methodology – leading to the development of a prototype system – for detecting fraudulent activities within FDBs. Their methodology (Figure 3.3) aimed to highlight potentially irregular activities arising on FDBs by looking at both comments posted on the FDB, and the share prices of companies.

This FMMS proposed by P. S. Lee, Owda, and Crockett (ibid.) considered textual comments collected from three different FDBs over a 12-week period, in which over 500,000 comments and 29 million stock prices were collected. The detection

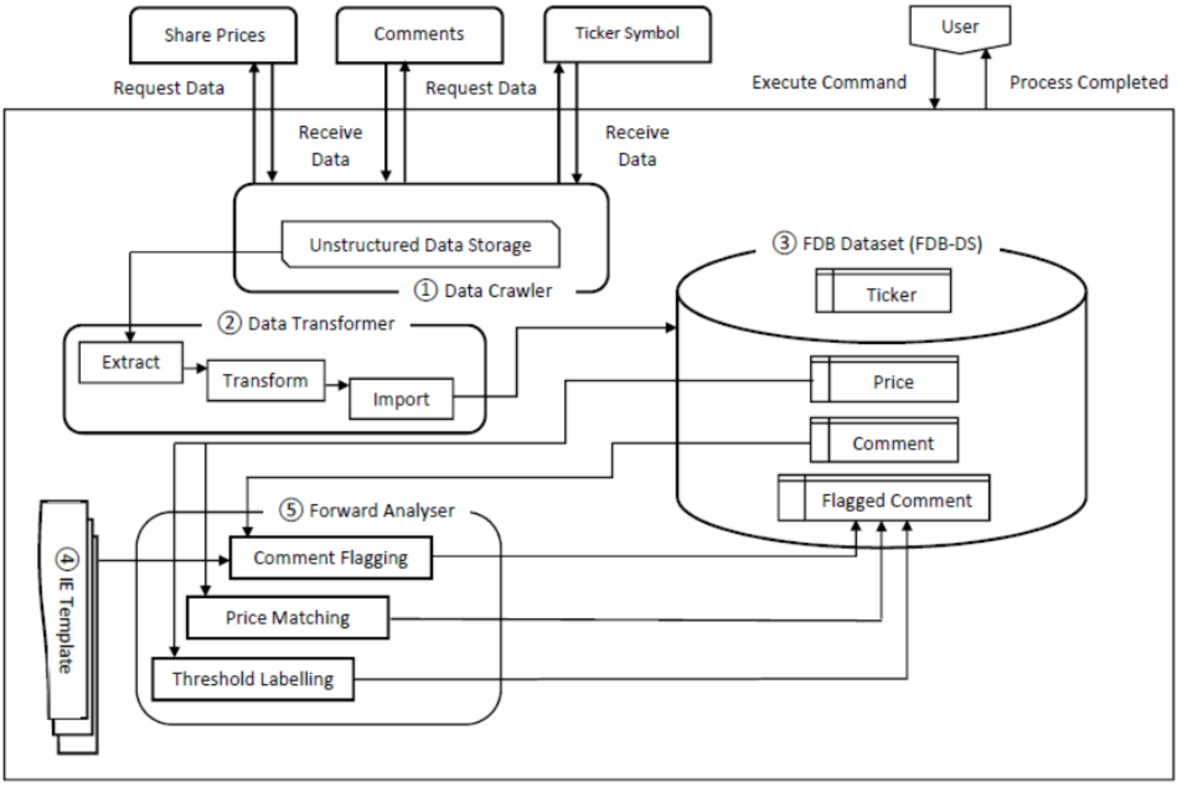


FIGURE 3.3: Architectural overview of an FMMS proposed by P. S. Lee, Owda, and Crockett (2018)

of irregular comments was achieved by creating a list of keywords associated with the common pump-and-dump manipulation technique. The presence of these keywords formed the basis of establishing irregular posts - which could ultimately be shown to an expert for consideration and further analyses. Determining the significance of the irregular posts followed a rule-based approach, whereby if the stock price of a company changed by a pre-defined threshold within two days, a label would be assigned highlighting the severity of the irregular post. These price hike thresholds assigned labels of red, amber, yellow - highlighting the severity of the flagged comment - and labels of C and N were used to denote a comment was not a cause for concern, or there was missing price data respectively. The results of applying this methodology to the comments collected resulted in 7.25% of comments assigned an R (red) label, 5.12% being assigned an A (amber) label, and 10.42% were assigned a Y (yellow) label. Over the two-week period, an average of 593 comments were flagged every day as either R, A, or Y - indicating the prominence of potentially irregular posts circulating on FDBs. As this FMMS primarily relied on a keyword list and focused on price movements over a two-day period, some potentially irregular comments could have gone undetected if they were not accompanied by significant price movements to meet the required thresholds, or terminology within the post was not considered in the keyword list. Any system which is heavily dependent on the presence of keywords is likely to need frequent maintenance of the lexicon to ensure new terminology used by investors is reflected within the lexicon.

3.5.2 Models Supporting Monitoring Systems

No FMMS operates independently - they are supported by various modules, models, and methodologies, each of which may contribute to the FMMS in different ways - such as data collection, storage, analysis and visualisation. Statistical and machine learning models have played an important role in the detection of irregular and fraudulent stock market activity. Several of these will now be examined.

Contextual-based irregularity detection has enjoyed success in detecting market manipulation in stock markets. Golmohammadi et al. (2015) designed and implemented a set of experiments to evaluate their proposed contextual irregularity detection model for time series stock data. The model considered not only the context of a time series in a specific time window but also considered the context of other time series in a similar group (e.g. two companies in the same sector or industry). Their experiments concluded that the proposed method could outperform existing approaches such as k-Nearest Neighbours and Random Walk in identifying time series grouped by company sector.

Y. Kim and Sohn (2012) developed a method to detect suspicious patterns of stock price manipulation using an unsupervised data mining technique known as peer group analysis. The developed model compares time-series stock prices of a company with other stocks that exhibit a similar pattern of price change and examined suspicious cases of stock manipulation using publicly available stock price data.

Clustering-based models have also been adopted to aid in the detection of irregularities within stock markets. Close and Kashef (2020) proposed combining an artificial immune system approach with clustering algorithms in order to detect potential irregular trading activity. The combination of using these two approaches allowed the models to adapt over time and adjust to normal trading behaviour as it evolves. The results of their study highlighted that their hybrid approach can be an effective tool for irregularity detection in the financial domain and is a competitive solution to the leading kernel density estimator approach that inspired their research.

3.6 Chapter Summary

This chapter has provided an overview of the literature of irregularity detection within financial markets, focusing on proposed FMMS and models that support the monitoring process.

Much of the existing literature is focused on analysing price movements, which naturally takes place after some irregular activity (e.g. spreading of rumours to inflate a share price). Little attention has been given to the behaviour of investors, such as the discussions taking place in different communicative environments and the volume of activity surroundings key company events. One of the key commonalities of the existing systems is their dependence on a plethora of data sources which include quantitative data such as share prices and share volume, and qualitative data such as investor discussion.

A critical shortcoming of existing FMMSs is that they are incredibly high-level and abstract, and lack the necessary implementation details for designing, developing, and deploying machine learning models to aid in the collection and analysis of such vast quantities of data. This research involves how such models can be combined to enrich a FMMS, including models that assist with the collection of tweets that contain naming conflicts (Section 6), assessing the credibility of such tweets that are financial in nature (Section 7), and how such data can be clustered to spot irregular activity (Section 8).

Evaluating FMMSs in a real-world scenario is challenging to undertake, as it involves comparing the results of a developed system against cases that were proven to be regarded as irregular or manipulative by a regulatory body. Regulatory authorities are also hesitant to give their stamp of approval to systems as it could signal the regulatory body's acceptance of a methodology or give manipulators insight into how irregular activity is detected on financial markets.

Chapter 4

Research Methodology

4.1 Overview

This chapter will discuss the research approach which was used to develop the Smart Data Ecosystem (SDE). Firstly, an overview of the SDE is provided, including the definition this thesis adopts of an ecosystem, is provided in Section 4.2.

4.2 Smart Data Ecosystem

As discussed in Section 1.3, the aim of this research is the creation of an ecosystem capable of detecting irregular behaviour. To this end, it is important to clarify that the aim is not to create an omniscient ecosystem capable of monitoring every discussion channel which exists - such a task would be impossible. Instead, the ecosystem will focus on selected discussion channels to provide a proof of concept, and be complemented by other data sources which will be introduced in this chapter.

4.2.1 What is a 'Smart' Data Ecosystem?

There have been several definitions relating to what an ecosystem is in the context of computing. The oldest, and original, definition of an ecosystem in a computing context is attributed to Messerschmitt and Szyperski (2005). Messerschmitt and Szyperski (ibid.) define an ecosystem as "a software ecosystem refers to a collection of software products that have some given degree of symbiotic relationships". The emphasis on symbiotic (a term originating in biology to refer to interaction between two different organisms living in close physical association) relationships is of particular interest in this research. Bosch and Bosch-Sijtsema (2010) refer to a software ecosystem as ″A software ecosystem consists of a software platform, a set of internal and external developers and a community of domain experts in service to a community of users that compose relevant solution elements to satisfy their needs″. This definition does not emphasise relationships between certain components, but instead focuses more so on cooperation between internal and external developers. Manikas and Hansen (2013) defines a software ecosystem as the ″interaction of a set of actors on top of a common technological platform that results in a number of software solutions or services″. This term refers to interaction between actors – the Unified Modelling Language (UML) terminology - to denote a ″role played by a user or any other system that interacts with the subject″ (Fowler 2004). Actors in this sense could represent a human actor or another system entirely. These definitions have motivated the definition of the ecosystem to be developed as part of this project, which will be introduced at the end of this section.

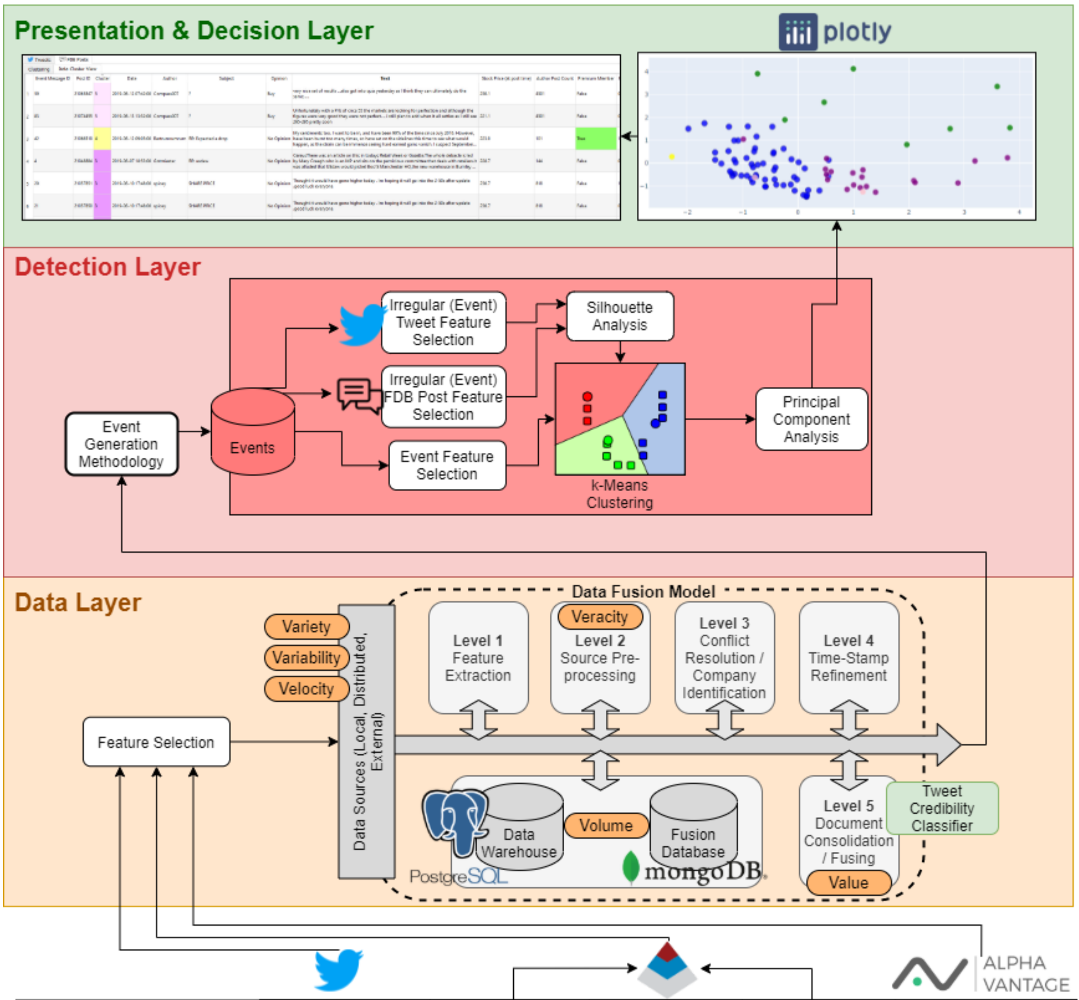
One of the principal concerns of any ecosystem is data management - how is data collected, cleaned, stored, and retrieved before analysis is undertaken on such data? The popularisation of conceptual big data models have detailed the prominent Vs of big data and the challenges associated with each. The seminal big data model proposed by Laney (2001) proposed 3Vs - volume, velocity, and variety. Extensions to this conceptual model have been proposed and adopted since then, which typically add on more Vs - with recent research by Khan et al. (2019) positing 51Vs of big data.

With the term ecosystem now defined, the *Smart* aspect needs to be addressed. *Smart* data is an organised way to semantically compile, manipulate, correlate, and analyse diverse data sources to get the most valuable V from the data - its *value* (Duong, Nguyen, and Jo 2017).

Based on the definitions of software ecosystems, and the consideration of the various Vs of big data, this thesis defines a smart data ecosystem (SDE) as **a series of cooperating layers to deal with the collection, cleaning, storage, and analysis of big data and tools to aid the visualisation process**. Aiding the visualisation process includes providing mechanisms in which a user can visualise the different clusters (groups) of tweets, FDB posts, and daily activity for a certain stock-listed company.

4.2.2 Smart Data Ecosystem Overview

The SDE (Figure 4.1) developed as part of this research is composed of three cooperating layers (Data, Detection, Presentation & Decision), each responsible for certain functions. Each of these layers will now be briefly outlined, with detailed explanations reserved for the chapters which correspond to the respective layers.


Presentation & Decision Layer
plotly
Detection Layer
Irregular (Event) Tweet Feature Selection
Irregular (Event) FDB Post Feature Selection
Event Feature Selection
Silhouette Analysis
Event Generation Methodology
Events
k-Means Clustering
Principal Component Analysis
Data Layer
Variety
Variability
Velocity
Data Sources (Local, Distributed, External)
Feature Selection
Data Fusion Model
Level 1 Feature Extraction
Veracity
Level 2 Source Pre-processing
Level 3 Conflict Resolution / Company Identification
Level 4 Time-Stamp Refinement
Data Warehouse
PostgreSQL
Volume
Fusion Database
mongoDB
Level 5 Document Consolidation / Fusing
Value
Tweet Credibility Classifier
ALPHA VANTAGE

Data Source	Twitter	London South East Forum	London South East Financial Diary Dates	London South East Broker Agency Pages	AlphaVantage
Data	Tweets	Financial Discussion Board Posts	Financial Diary Dates	Broker Agency Ratings	Daily & Intraday Share Prices
Collection Technique	Tweepy API	Scrapy Web Crawler	Scrapy Web Crawler	Scrapy Web Crawler	AlphaVantage API

FIGURE 4.1: Smart Data Ecosystem Diagram

4.2.3 Data Layer

The foundational layer of the SDE is the data layer. The data layer is responsible for collecting data from a variety of sources - which is the primary input to this layer. Chapter 5 will provide a detailed explanation and justification for these data sources, including details of fusing these heterogeneous data to assist the proceeding layers. This layer also deals with a crucial step in analysing Twitter stock discussion - the resolution of tweets containing cashtag collisions (Chapter 6). The credibility of financial stock tweets is also undertaken at this layer, using the novel methodology presented in Chapter 7. The output of this layer is a time-slice window covering two dimensions: (1) the company to which the data refers to, and (2) the date.

The outputs of the data layer are two-fold (1) company-specific time-window documents that contain all of the data pertinent to that window for that company and (2) event-based documents that contain all data pertinent to a company event.

4.2.4 Detection Layer

Responsible for the detection of irregularities, the detection layer (Chapter 8) is concerned with looking at time-slice windows and events provided by the data layer. The detection layer makes use of the popular unsupervised k-means clustering algorithm to identify new patterns of posting behaviour and detecting irregularities surrounding *events*. This thesis defines an event in the financial context as a **"moment of significance in a company's operations"** - this could include a new Chief Operating Officer being appointed, or a broker agency offering a favourable/unfavourable analyst rating for the company.

The output of the detection layer are the results of performing k-means clustering on a specific event and/or - these results are then fed to the presentation & decision layer for visualisation and further analysis to support the decision-making process.

4.2.5 Presentation & Decision Layer

The final layer of the ecosystem deals with the presentation and decision-making elements of the ecosystem. The clustering output from the previous layer is visualised by adopting a dimensionality-reduction algorithm (principal component analysis) to allow easier interpretation of the clustering output. Various tools and visualisations are provided in this layer, assisting in the decision-making process and establishing if any detected irregularities warrant further investigation.

4.2.6 SDE Companies

Throughout this thesis, reference will be made to companies in which the SDE actively collects data for. The full list of companies can be found in Appendix B. These companies were selected based on the following criteria:

- Companies were first selected from each industry
- Companies must have been listed on the LSE for at least two years (to maximise the chances of data collection)

4.3 Chapter Summary

This chapter has provided a high-level overview of the SDE that has been developed as part of this research. Each of the layers and their responsibilities have been defined, along with their respective inputs and outputs. The next chapter will introduce the first layer of the ecosystem: the data layer, which will provide details of data collection, storage, and pre-processing steps carried out on each of the data sources. The tools which are developed across these layers are evaluated with the assistance of financial market experts through qualitative interviews (Chapter 9).

Chapter 5

Data Layer: Foundations & Data Fusion Model

5.1 Overview

This purpose of this chapter is to provide a detailed overview of the data layer of the Smart Data Ecosystem (SDE) that was formally introduced in Section 4.2. This layer features several important aspects of the ecosystem that will be explored in this chapter, Chapter 6, and Chapter 7.

The ineluctable growth of heterogeneous stock market data poses a serious challenge to regulators and researchers that attempt to analyse stock market prices and discussion for purposes such as predicting stock prices and monitoring for irregular behaviour (Flood, Jagadish, and D 2016; Ngai et al. 2017). Stock time-series data is typically published in various frequencies that include minutely, hourly, and daily intervals. Such time-series data includes the open, high, low, and close prices of the stock during the given interval, including the volume of shares traded within the window. Alongside this structured stream of numeric data is the discussion taking place of those stocks by the investors. These investors have a wide range of platforms to use to discuss and disseminate information on such stocks, ranging from Online Social Networks (OSNs) which include the Twitter microblogging site, StockTwits, and numerous Financial Discussion Boards (FDBs). How can such data be combined, and what are the advantages of performing such a task? The aim of this chapter is to answer such questions.

This chapter firstly presents a definition of data fusion (Section 5.2), followed by a review of popular models for performing data fusion (Section 5.3). Then, an overview of the literature on data fusion, specifically within the domain of stock markets (Section 5.5), is presented. The data sources utilised by the SDE are then introduced (Section 5.6). The data fusion model utilised by the data layer of the SDE (the principal contribution of the data layer, introduced in Section 4) is then provided (Section 5.7). This chapter is motivated by the research undertaken and published in Evans, Owda, Crockett, and Ana Fernández Vilas (2018) (Appendix A).

5.2 Data Fusion

One of the most well-known definitions of data fusion was provided by Hall and Llinas (1998), in which they defined data fusion as *"data fusion techniques combine data from multiple sensors and related information from associated databases to achieve improved accuracy and more specific inferences than could be achieved by the use of a single sensor alone"*. Data fusion has become a firmly established practice for handling heterogeneous data sources by associating and combining data sources together (Alyannezhadi, Pouyan, and Abolghasemi 2017; Bleiholder and Naumann 2008). The use of such techniques can be seen as a systemic approach - whereby the whole is bigger than the sum of its parts - and relying on single sensors in isolation of themselves does not provide much value - it is when such data is combined in some way there the value is unlocked.

Data fusion is utilised in many fields, including healthcare (Y. D. Zhang et al. 2020; Shen et al. 2021; Qi et al. 2020), internet of things (Ullah and Youn 2020; Aldeco-Pérez and Moreau 2008), and network intrusion (G. Li et al. 2018). The data fusion process is an incredibly domain-dependent task, meaning one approach may enjoy success in one domain but fail in another (Bleiholder and Naumann 2008).

5.3 Existing Data Fusion Models

The process of combining multiple data sources (often from multiple streams or sensors) is a process that has been well-documented and refined over the years. This

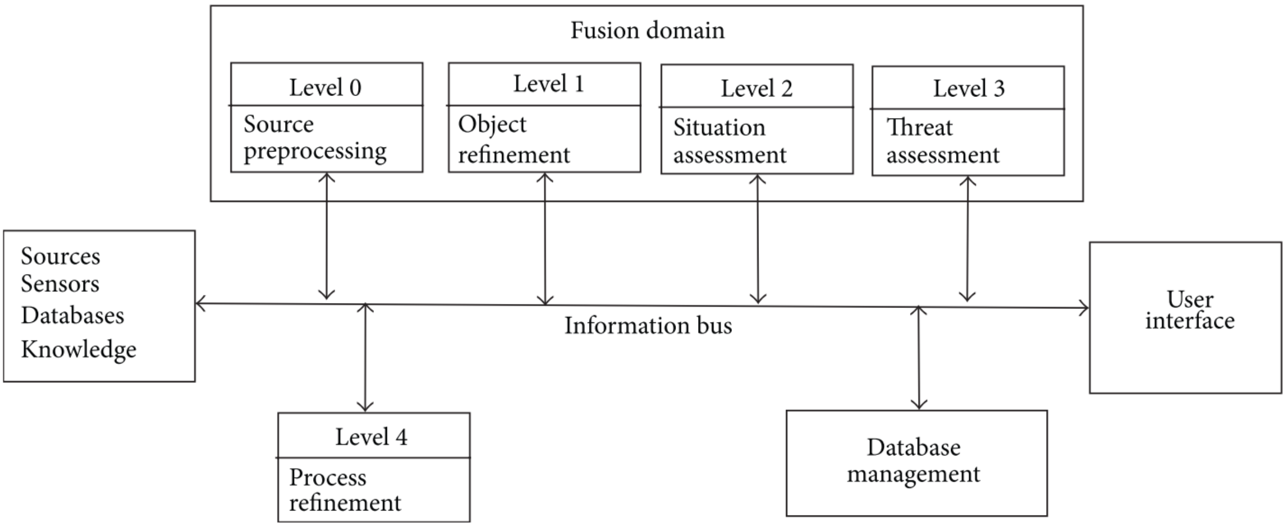


FIGURE 5.1: Joint Directors of Laboratories (JDL) Model

section will review the most prominent models and architectures proposed to carry out data fusion and the benefits and limitations of such models.

5.3.1 Joint Directors of Laboratories (JDL) Model

The Joint Directors of Laboratories fusion model (Figure 5.1) was first proposed by the US Department of Defense (DoD) in 1986 and is widely considered to be the seminal model for modelling the data fusion process (Blasch et al. 2013). Naturally, as this model was proposed by the DoD, its use case was intended for military applications such as battlefield surveillance, control of autonomous vehicles, and automated target recognition (Hall and Llinas 1998).

The elements of the JDL model are as follows:

- **Data Sources** - Sources of data to be fused would include sensor data (e.g. movement/weather), databases, a priori information references or geographic data, and human inputs (knowledge).
- **Level 0 - Source Pre-processing** - This level aims to reduce the volume of data by utilising data cleaning techniques, addressing missing values, and maintaining valuable information for the higher-level processes.
- **Level 1 - Object Refinement** - The object refinement level makes use of the processed data from the previous level. Common processes at this level include

Spatio-temporal alignment, state estimation, and the removal of false positives (McDaniel 2001).

- **Level 2 - Situation Assessment** - This level attempts to identify the likely situations based on the observed events and data that has been obtained. The output of this level is a group of high-level inferences.
- **Level 3 - Threat Assessment** - The purpose of this level is two-fold: (1) to evaluate the risk (or threat) and (2) predict the most logical outcome.
- **Level 4 - Process Refinement** - This level monitors system performance and involves handling real-time constraints.
- **Database Management** - At the commencement of the previous data fusion levels, the database management system stores the fused results.
- **User Interface** - The final aspect of the JDL model encompasses the human-computer interaction element of the data fusion process. Once data is fused, it is often used by a human operator in some way, such as undertaking analysis of the fused data or for visualisation purposes.

One of the key limitations of the JDL model is the uncertainty surrounding how previous or subsequent results could be utilised to further enhance the fusion process (feedback loop) (Castanedo 2013). Researchers, such as Meng et al. (2020), however, have noted that although the JDL model was primarily aimed at military applications, it is relative easy to adapt to other domains.

5.3.2 Dasarathy Fusion Model

Another popular data fusion model is the hierarchical Dasarathy model (Figure 5.2) (Dasarathy 1997). The Dasarathy model involves three levels of abstraction within data fusion: (1) data, (2) features, and (3) decisions. Dasarathy (ibid.) noted that data fusion could be done in and across all three of these abstract levels.

This model categorises the process of data fusion into six distinct categories (each of which incorporates the previously mentioned three abstract levels):

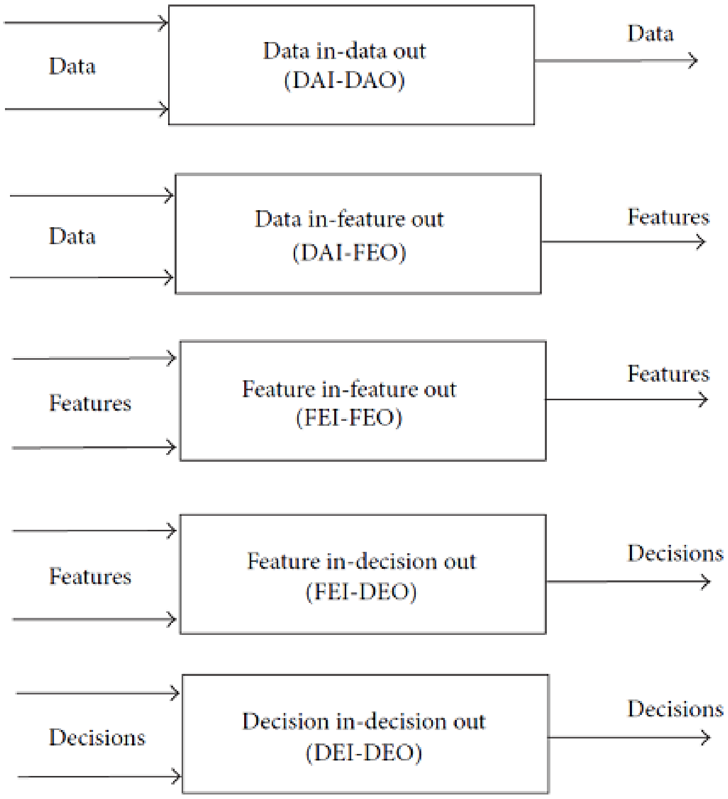


FIGURE 5.2: Dasarathy's fusion model (Dasarathy 1997)

1. **Data in - Data out (DAI-DAO)** - The first category is the most common type, in which both the input and output is typically low-level raw data obtained directly from a sensor (source). This level is often synonymous with the general term *data fusion*.

2. **Data in - Feature out (FAI-FEO)** - The second category fuses data input into a feature output. This will typically involve data from multiple sensors (sources) to generate a feature that is more informative than the individual data points themselves.

3. **Data in - Decision out (DAI-DEO)** - The third group involves raw data as the input and a decision as the output. This category is similar to FEI-DEO and is relevant to pattern recognition problems.

4. **Feature in - Feature out (FEI-FEO)** - The fourth category involves both the input and output being features (each are combinations of data points), and is often referred to as simply *feature fusion*.

5. **Feature in - Decision out (FEI-DEO)** - The fifth category includes feature as the input, with a decision as the output. The most common type of such a fusion task is supervised learning, in which features - some of which may be engineered from combinations of data points - are used to predict (decide) on a class (category).

6. **Decision in - Decision out (DEI-DEO)** - The final category involves both the input and output being decisions. This fusion type is particularly appropriate where there exists a need to combine decisions from an array of sources where different tasks and configurations exist. An example of this could be taking the predictions (decisions) of multiple machine learning models, and taking the majority opinion, hence forming a new decision.

Naturally, given the abstract nature of the Dasarathy model, it is not necessarily a framework that allows for new fusion models to be created, but is more of a framework to allow models to be compared.

5.4 Data Fusion Challenges

The fusion of disparate data sources is not without its challenges. The main challenges associated with combining different data sources together were outlined by Khaleghi et al. (2013) as being:

- **Disparate data** - Data pertaining to stocks is ubiquitous and comes in many forms. Numerical stock prices and unstructured user-generated discussion (e.g. posts and tweets) are all structured differently and require different methods to collect and clean.

- **Timestamps** - An issue following on from disparate data sources is the issue of timestamps. Although many APIs will provide a detailed time-stamp of the data points, other collection methods, such as web scraping, may not. Timezone differences may also exist depending on the collection method, including daylight saving times in some timezones.

- **Out-of-sequence data** - Leading on from the issue of individual time-stamps is the issue of time-series data. Time-series data, such as stock prices over a given period, are organised as discrete pieces of data, each labelled with a timestamp that aligns the data point to a specific point in time. Data points which are missing such time-stamps cannot be reliably fused with other data sources.

- **Conflicting data** - There are many APIs to choose from when collecting data such as stock prices. If several such APIs report different stock prices for a stock simultaneously, which of those is correct? An important aspect of any data fusion model is not to simply discard such data points, but provide a means of cross-checking such data points to ensure correctness.
- **Outliers** - Noise and outliers within datasets pose a significant issues in the fusion process. A data fusion model should provide some means to handle such imperfections, such as highlighting such outliers before concluding the fusion process.

5.5 Data Fusion in the Stock Market

The use of data fusion in the stock market is primarily aimed towards aiding stock market price prediction (Thakkar and Chaudhari 2021), a research area that is also popular in the field of machine learning. This section will explore the successes of data fusion within the context of the stock markets, including stock price prediction (Weng, M. A. Ahmed, and Megahed 2017; X. Zhang et al. 2018) and risk/return forecasting (L. Zhang et al. 2013) which is relevant to the work undertaken in this thesis.

5.5.1 Stock price prediction through data fusion

Dominating the literature on stock market data fusion, stock market price prediction has been a fast-growing field amongst researchers (Thakkar and Chaudhari 2021).

Weng, M. A. Ahmed, and Megahed (2017) proposed predicting one-day-ahead stock price movement by combining crowd-source knowledge bases (Google and Wikipedia platforms) with historic stock market data to establish if utilising such crowd-source data streams could lead to more accurate price predictions. A financial expert system (Figure 5.3) was developed in which a ”knowledge base” was scraped and formed from four different data sets: (1) historical stock market data, (2) commonly used technical indicators, (3) Wikipedia traffic statistics relating to a stock company’s page (e.g. general company profile, stock page, and pages relating to the company’s main products and services) and (4) Google News. The case study used to validate their methodology showed that the addition of online sources (Google and Wikipedia hits) gave better predictive power (85.8%) than the price and technical indicators alone (61.6%). However, this work only considered the stock price of Apple (NASDAQ:APPL) over a single time period, meaning the system may have generalised to large-cap stocks in which a wealth of data (Wikipedia and Google news stats) is available which may not necessarily be true of smaller-cap stocks. In addition, the authors noted that they expected a diminishing return with the inclusion of new data streams.

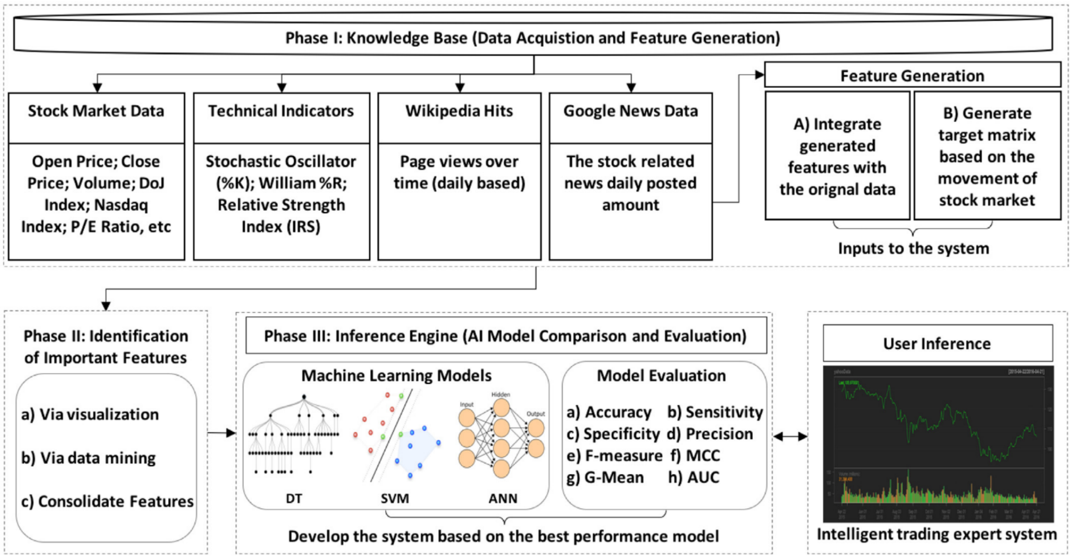

Phase I: Knowledge Base (Data Acquistion and Feature Generation)
Stock Market Data
Open Price; Close Price; Volume; DoJ Index; Nasdaq Index; P/E Ratio, etc
Technical Indicators
Stochastic Oscillator (%K); William %R; Relative Strength Index (IRS)
Wikipedia Hits
Page views over time (daily based)
Google News Data
The stock related news daily posted amount
Feature Generation
A) Integrate generated features with the orignal data
B) Generate target matrix based on the movement of stock market
Inputs to the system
Phase II: Identification of Important Features
a) Via visualization
b) Via data mining
c) Consolidate Features
Phase III: Inference Engine (AI Model Comparison and Evaluation)
Machine Learning Models
DT
SVM
ANN
Model Evaluation
a) Accuracy b) Sensitivity
c) Specificity d) Precision
e) F-measure f) MCC
g) G-Mean h) AUC
Develop the system based on the best performance model
User Inference
Intelligent trading expert system

FIGURE 5.3: Stock price prediction framework developed by Weng, M. A. Ahmed, and Megahed (2017)

X. Zhang et al. (2018) have attempted to leverage crowd-source information for the purpose of price prediction. The system developed by X. Zhang et al. (ibid.) (Figure 5.4) leverages events, sentiments, and qualitative features extracted from sources including web news, social media, and quantitative stock prices. Events were extracted from web news articles, along with user sentiments collected from social media to investigate their joint impacts on the movements of stock prices. A tensor was contributed to fuse the heterogeneous data and capture the intrinsic relationship between the events and the sentiments of the investors. A case study involving the companies listed on the Chinese stock exchange (China A-Share) was conducted to demonstrate the effectiveness of the model. When utilising the additional crowd-sourced data sources, the developed model was able to outperform models that only took into account the quantitative stock data. The authors did note that such a model is limited by not adopting advanced natural language processing techniques, which could be included to learn event presentation by incorporating domain knowledge and better categorisation of events.

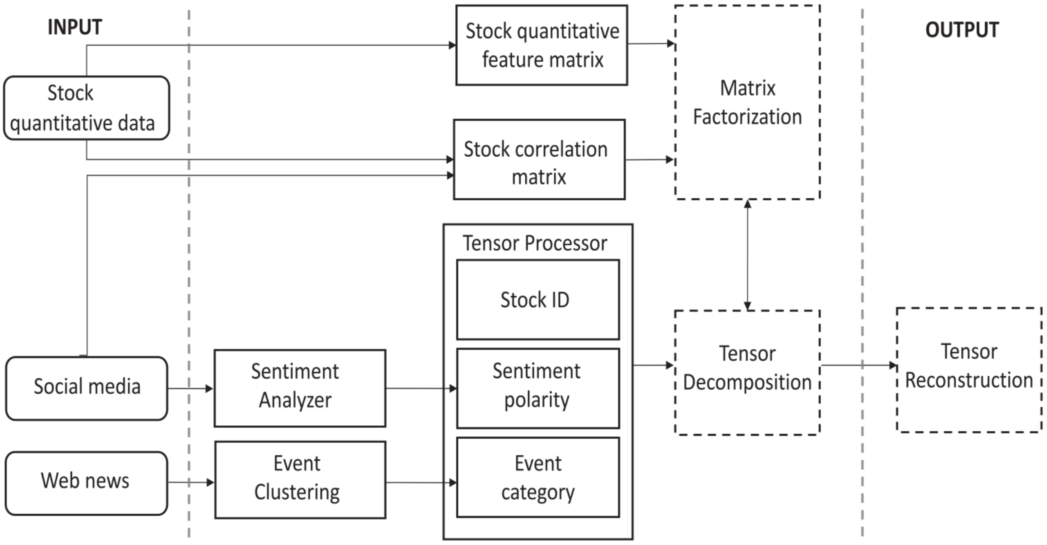


FIGURE 5.4: Stock price prediction framework developed by X. Zhang et al. (2018)

5.5.2 Risk/return forecasting

One of the principal concerns of an investor is to reduce financial loss as much as possible. To this end, an early-warning system to warn investors that stock prices may begin to fall has become a point of interest amongst researchers (Thakkar and Chaudhari 2021; L. Zhang et al. 2013).

L. Zhang et al. (2013) proposed an early-warning system that predicts potential stock price decline, which is enhanced by data fusion. This system adopted the Dempster-Shafter theory - a general extension of Bayesian theory - that fused 25 independent features together to derive the likelihood of financial loss from stock price decline. These features included many values derived from fundamental stock analysis (e.g. quick ratio, liquidity ratio, earning per share).

Stock market trends can be visualised using historical stock price data. In the same vein, market news can reflect different events, their impact on stock prices (Schumaker and H. Chen 2006; X. Li, Xie, et al. 2014; Q. Li et al. 2014), discussion between investors on corresponding analysis, in addition to market behaviours (Thakkar and Chaudhari 2021). X. Li, X. Huang, et al. (2014) proposed an integrated approach that applied information fusion of market news and stock prices to predict intra-day stock return.

5.5.3 Summary of related work

Much of the related work on data fusion within financial market is too focused on stock price prediction and has been neglected in other applications (e.g. irregularity and fraud detection). The fusion of data from OSNs (e.g. tweets) with other sources (e.g. financial discussion boards and quantitative stock prices) has also not been explored or exploited in the context of financial market monitoring. The challenges of data fusion discussed in Section 5.4 also need to be addressed, with Section 5.7 attempting to shine some light on how this can be achieved when the SDEs data fusion model is presented. The data fusion model of the SDE (Section 5.7) will address these shortcomings, providing mechanisms in which data from a variety of channels can be combined to aid in the detection of financial market irregularities.

5.6 Data Layer: Data Feeders & Collection

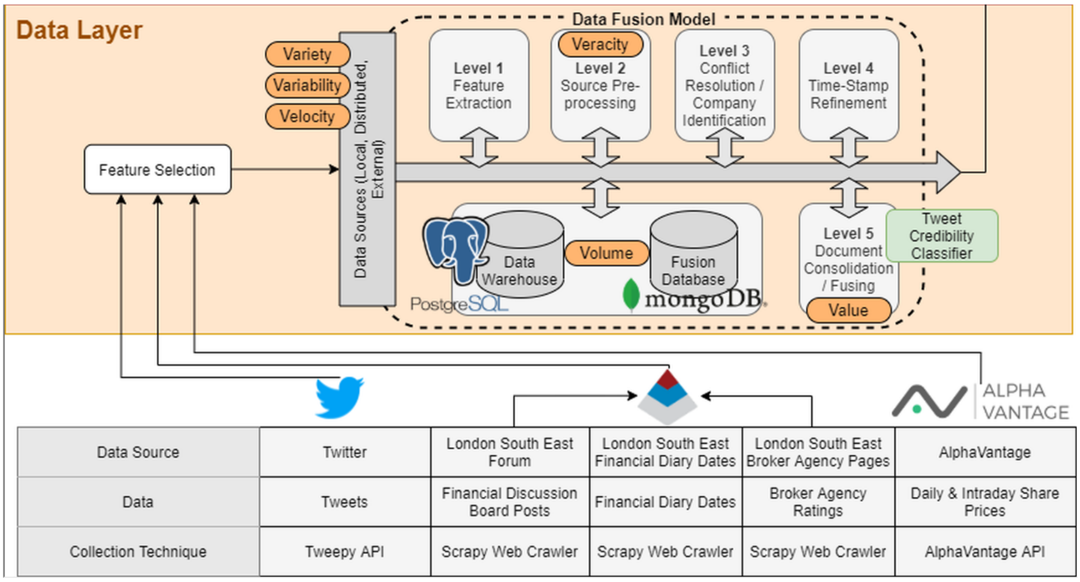
With the related work on stock market data fusion now explored, the proposed SDE's data fusion model will be presented in Section 5.7. Data fusion is the main novelty of the data layer, in which data sources are combined to benefit other tasks, such as the detection of irregularities, and provide mechanisms in which irrelevant data is discarded. Before this, however, the data sources the SDE considers will be introduced, along with the mechanisms used to collect them. A high-level overview of these data sources and collection techniques is presented in Table 5.1. All mentions of data collection in this section refer to data collected for 200 shortlisted companies

TABLE 5.1: SDE Data Sources

Data Source	Data	Collection Mechanism
Twitter	Tweets	Tweepy[1] (Twitter Streaming API)
London South East	Financial Discussion Board Posts Financial Diary Dates Broker Ratings	Scrapy[2]
AlphaVantage	Intraday Share Prices (15-min intervals) Daily Share Prices	AlphaVantage[3] API

(Appendix B) from the London Stock Exchange (LSE). Recall from Section 2.3 that the LSE has over 2,000 companies listed at any one time, meaning the SDE actively collects data for around 10% of those companies. The companies were shortlisted based on their industry, and must have been listed on the LSE for at least two years (to maximise the chances of data being collected).

The data layer (Figure 5.5) of the SDE (introduced in Chapter 4) is composed of data feeders, each responsible for collecting data from a specific service or website. Several of the data sources considered by the SDE feature APIs to streamline the collection from such sources. However, some data sources do not possess mechanisms to collect structured data, meaning web scraping techniques will need to be adopted to collect such data. Once the data is collected, it is fused into time-slice windows to provide synchronicity for the different data sources (discussed in Section 5.7).


Data Layer
Variety
Variability
Velocity
Data Sources (Local, Distributed, External)
Feature Selection
Data Fusion Model
Level 1 Feature Extraction
Veracity
Level 2 Source Pre-processing
Level 3 Conflict Resolution / Company Identification
Level 4 Time-Stamp Refinement
Data Warehouse
PostgreSQL
Volume
Fusion Database
mongoDB
Level 5 Document Consolidation / Fusing
Value
Tweet Credibility Classifier
ALPHA VANTAGE

Data Source	Twitter	London South East Forum	London South East Financial Diary Dates	London South East Broker Agency Pages	AlphaVantage
Data	Tweets	Financial Discussion Board Posts	Financial Diary Dates	Broker Agency Ratings	Daily & Intraday Share Prices
Collection Technique	Tweepy API	Scrapy Web Crawler	Scrapy Web Crawler	Scrapy Web Crawler	AlphaVantage API

FIGURE 5.5: Data layer of the SDE

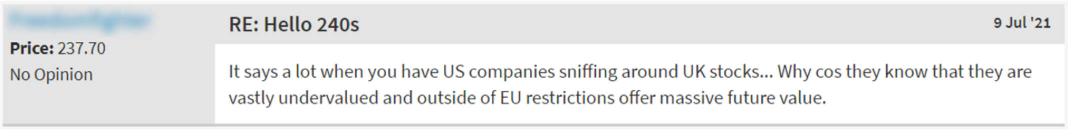
5.6.1 Tweet Collection

Tweets are collected in real-time by the SDE via Tweepy[4], which is a wrapper to Twitter's streaming API. This API collects approximately 1% of all tweets in real-time (K. Chen, Duan, and S. Yang 2021), and returns such tweets as a JavaScript object notation (JSON) object. All tweets (along with the metadata within) collected by the SDE are initially stored in a data warehouse prior to being combined with other data sources (Section 5.7).

5.6.2 London South East Data Collection

As discussed in Section 2.5.1, London South East[5] is a popular website that features an FDB for the discussion of stocks listed on the London Stock Exchange, including aggregating other company-specific information such as broker ratings and key financial diary dates.

FDB posts available on London South East have various metadata associated with them, some of which require an active and logged in London South East account to view (Summarised in Table 5.2). An example post based on viewing a forum page while not logged in is shown in Figure 5.6, with the same post shown in Figure 5.7 with an account logged in. All of the attributes shown in Table 5.2 are collected from posts, using a Scrapy spider capable of logging into the London South East website.

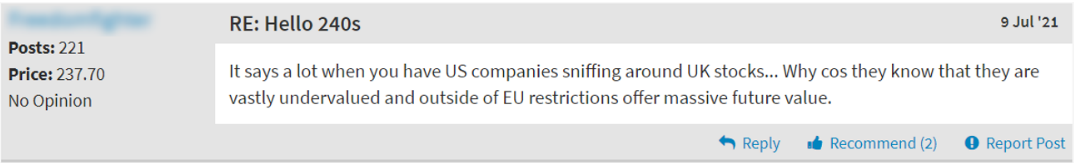


[redacted]
Price: 237.70
No Opinion

RE: Hello 240s 9 Jul '21

It says a lot when you have US companies sniffing around UK stocks... Why cos they know that they are vastly undervalued and outside of EU restrictions offer massive future value.

FIGURE 5.6: An example London South East post (not logged in)



[redacted]
Posts: 221
Price: 237.70
No Opinion

RE: Hello 240s 9 Jul '21

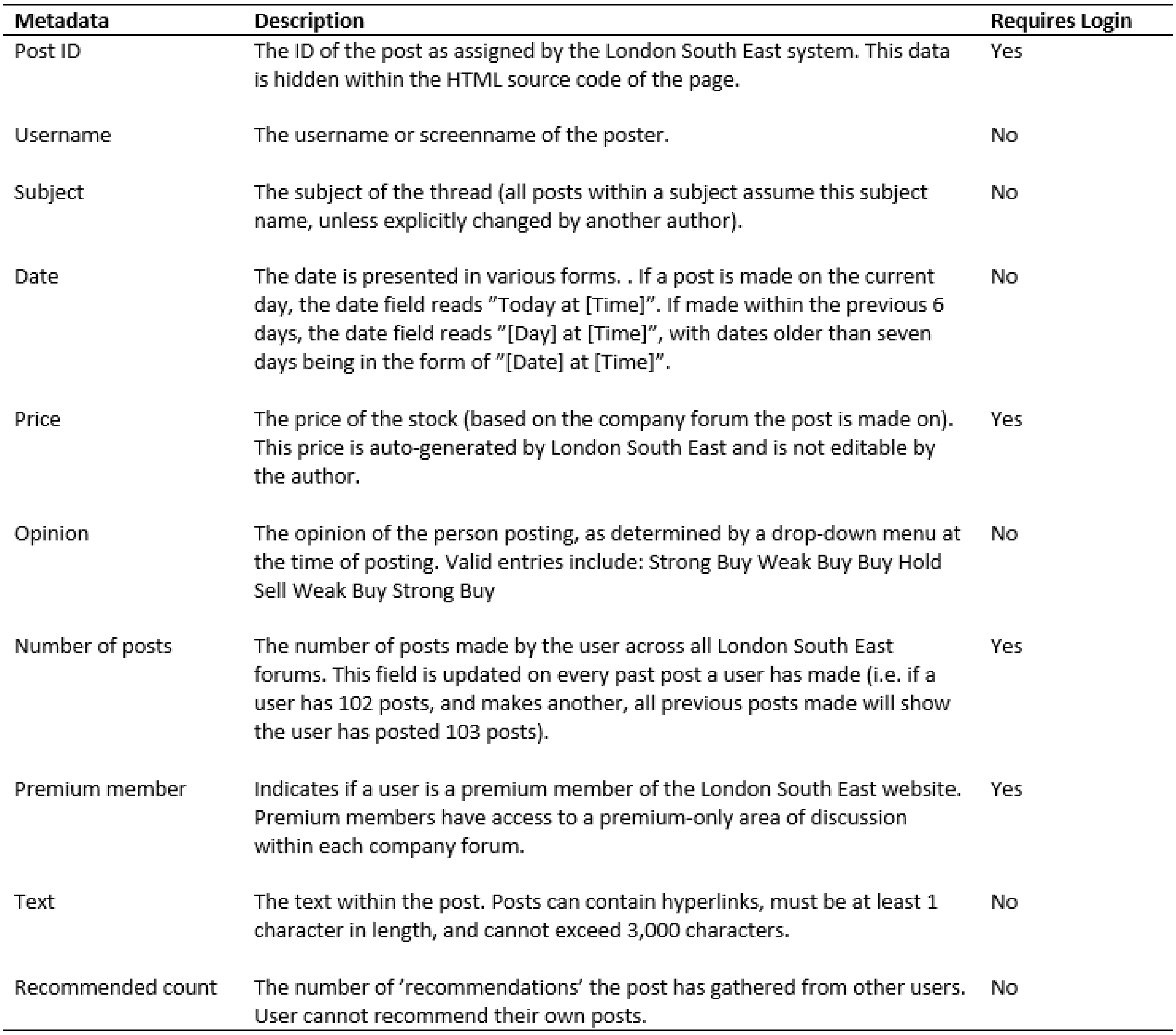
It says a lot when you have US companies sniffing around UK stocks... Why cos they know that they are vastly undervalued and outside of EU restrictions offer massive future value.

FIGURE 5.7: An example London South East post (logged in)

[4] https://www.tweepy.org/
[5] https://www.lse.co.uk/

Financial diary dates (FDDs) are dates that hold some significance to a company. These include dates in which dividend payments are made to investors, trading announcements, and dates of annual and emergency general meetings. All FDDs located on each company's FDD page on London South East are scraped and stored within the SDE.

TABLE 5.2: London South East FDB post attributes



Metadata	Description	Requires Login
Post ID	The ID of the post as assigned by the London South East system. This data is hidden within the HTML source code of the page.	Yes
Username	The username or screenname of the poster.	No
Subject	The subject of the thread (all posts within a subject assume this subject name, unless explicitly changed by another author).	No
Date	The date is presented in various forms. . If a post is made on the current day, the date field reads "Today at [Time]". If made within the previous 6 days, the date field reads "[Day] at [Time]", with dates older than seven days being in the form of "[Date] at [Time]".	No
Price	The price of the stock (based on the company forum the post is made on). This price is auto-generated by London South East and is not editable by the author.	Yes
Opinion	The opinion of the person posting, as determined by a drop-down menu at the time of posting. Valid entries include: Strong Buy Weak Buy Buy Hold Sell Weak Buy Strong Buy	No
Number of posts	The number of posts made by the user across all London South East forums. This field is updated on every past post a user has made (i.e. if a user has 102 posts, and makes another, all previous posts made will show the user has posted 103 posts).	Yes
Premium member	Indicates if a user is a premium member of the London South East website. Premium members have access to a premium-only area of discussion within each company forum.	Yes
Text	The text within the post. Posts can contain hyperlinks, must be at least 1 character in length, and cannot exceed 3,000 characters.	No
Recommended count	The number of 'recommendations' the post has gathered from other users. User cannot recommend their own posts.	No

As is the case with many discussion board sites, APIs to collect user posts are often not available. For this reason, the Scrapy[6] web scraping framework is utilised to crawl pages. Scrapy is a high-level web crawling framework that can be used to crawl websites and extract data from web pages. When collecting information

[6] https://scrapy.org/

using Scrapy, this collection process only occurred during off-peak hours (when the LSE was not open for trading), and the scraping process was throttled in order to minimise any disruption to their services.

5.6.3 Share Price Collection

AlphaVantage[7] provides market data ranging from traditional asset classes (e.g. stocks and exchange-traded funds), to forex and cryptocurrency data. Stock prices are available in both daily and intraday intervals using the popular open, high, low, close (OHLC) variants, in addition to the amount of shares traded in a given window (volume). Daily OHLC prices and intraday OHLC prices in 15-minute intervals are collected for all of the shortlisted SDE companies (Appendix B).

5.7 Data Layer: Data Fusion Model

This section will introduce the Data Fusion Model (DFM) which will be developed and utilised by the SDE introduced in Section 4.2 (Figure 5.8). The previously discussed JDL model (Section 5.3.1) inspires the data fusion model. Data is fused based on two dimensions: (1) the company the data refers to, and (2) the date.

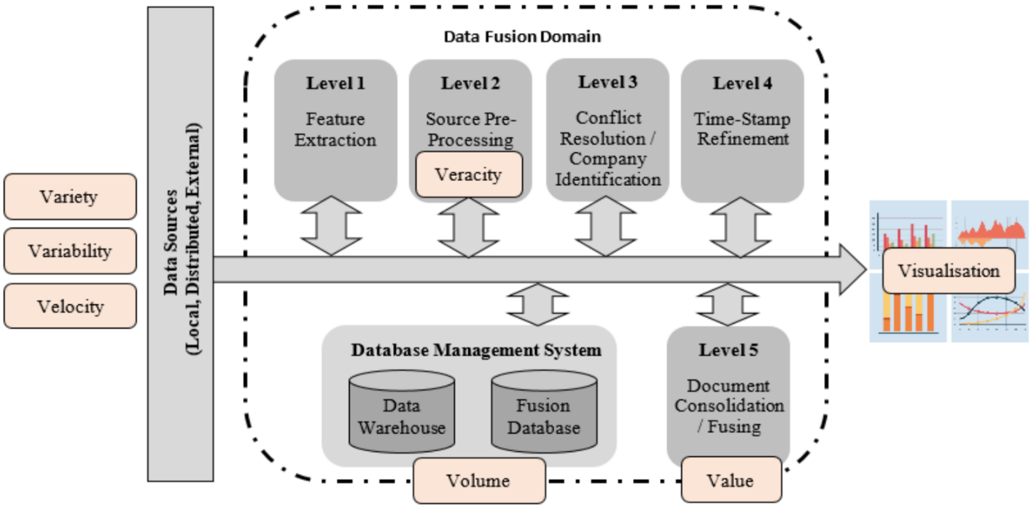


FIGURE 5.8: SDE Data fusion model

The default configuration of the data fusion model is to fuse data into daily time-slice company windows. A day is a significant time unit within the stock market;

[7] https://www.alphavantage.co/

stock prices and indeed stock indices are often summarised by their daily performance. Much of the existing research which attempts to use or predict stock prices uses this time-window. From a trading perspective, a day is an important time horizon: day traders are required to close their portfolios at the end of the day, and long-term strategies (e.g. large investment portfolios) are often adjusted daily (Eisler, Kertész, and Lillo 2007).

The DFM (Figure 5.8) is made up of the following elements:

- **Data Warehouse** - The data warehouse serves as a repository for all data the SDE will use, prior to any fusion taking place.
- **Level 1 - Feature Extraction** - The first level in the DFM involves the extraction of features from the data sources. Naturally, not all available features of a data source will have value from being combined with other sources. Therefore, any redundant features are discarded at beginning of the fusion process.
- **Level 2 - Source Pre-processing** - The second level deals with common pre-processing steps such as data cleaning, normalisation, missing value imputation, identifying noise and outliers, and transformation. Techniques such as Named Entity Recognition, Stemming, and Lemmatisation are adopted at this level to prepare the data for the machine learning process required later on in the ecosystem.
- **Level 3 - Conflict Resolution / Company Identification** - One of the key challenges of associating data sources with specific entities (e.g. companies), is that some data may not explicitly mention a company by name. This is a particularly cumbersome issue for social media channels such as Twitter, where investors will only include a company's ticker symbol (via the cashtag mechanism) and not the company name explicitly. The DFM addresses this issue at this layer, utilising the methodologies discussed in Chapter 6.
- **Level 4 - Timestamp Refinement** - Timestamps are a decisive factor as to whether data from multiple sources can be combined. APIs may return timestamps in local time, and specifically for the UK, may include daylight saving hours. This level refines timestamps to deal with such discrepancies.

- **Level 5 - Document Consolidation / Fusing** - Once timestamps have been addressed and aligned, the conclusion process concludes by grouping all of the related data in a fusion document within the fusion database. The fusion document is the name used to describe the data that has been fused together and stored in a NoSQL document - the term used to denote the grouping of data in key-value pairs.

- **Fusion Database** - The fusion databases houses all of the time-slice company windows. The SDE utilises MongoDB as its fusion database, which supports the popular document-oriented NoSQL model. A NoSQL database has been chosen as it is a document-oriented database system, meaning it will support the fusion document style approach previously discussed.

The DFM presented in this section was implemented and successfully collects, filters (in the case of tweets containing naming collisions), cleans, and stores data relating to the sources previously mentioned in Section 5.6.

5.8 Chapter Summary

This chapter has provided a background on data fusion, focusing primarily on stock market use cases. Naturally, the popularity of stock market price prediction means that a large body of work on data fusion in the stock market domain is largely aimed at this area. The DFM model to be incorporated within the SDE has also been presented, including the specific levels where issues relating to stock data (e.g. timestamp refinement and resolving cashtag collisions) are addressed.

The main contributions of this chapter is the novel SDE DFM for the fusing of disparate financial data sources. It is envisaged that combining multiple financial data sources together will aid in the detection of irregularities pertaining to stock discussion by constructing time-window and company-specific windows.

The next chapter will delve deeper into the data layer, and provide details on one of the key contributions of this work which is located at level 3 of the DFM: the resolution of cashtag collisions in stock tweets.

Chapter 6

Data Layer: Resolving Colliding Cashtags within Tweets

6.1 Overview

This chapter provides an overview of one of the most crucial tasks of the data layer of the SDE (introduced in Section 4.2): resolving cashtag collisions present in tweets. As described in Section 5.7, tweets can contain cashtags - clickable hyperlinks that mimic a stock's ticker symbol, prefixed with a $ symbol. The issue with searching and collecting such tweets, however, is that companies on different exchanges often possess identical ticker symbols. This research has coined the term 'cashtag collision' to refer to this phenomenon. Twitter does not currently attempt to resolve such conflicts, meaning it is left to the investor to decipher if a tweet relates to the company they are interested in discovering news for. One of the functionalities of the detection layer (Chapter 8) is to cluster tweets for a given time window to detect potentially irregular tweets. It is critical to resolve cashtag collisions before undertaking this clustering process so that tweets not relating to the LSE are discarded and do not adversely impact the clustering process.

The aim of this chapter is to provide an overview of two methodologies: (1) the creation of custom company corpora (Section 6.5), and (2) a methodology for resolving cashtag collisions (Section 6.6). These methodologies, and an experiment carried out to validate them, has been published in Evans, Owda, Crockett, and Ana Fernandez Vilas (2019) (Appendix C). This paper presented the related methodologies and an experiment involving 1,000 annotated tweets containing a cashtag referencing at least one of 100 LSE companies (listed in Appendix E). This chapter will adapt the methodology and involve an experiment on a larger dataset of 5,000 tweets, using the ticker symbols of the 200 SDE-shortlisted companies (Appendix B).

This chapter begins by providing an overview of cashtag collisions and the issues arising from this phenomenon (Section 6.2). Related work that has utilised the cashtag mechanism is discussed (Section 6.3), along with the issues such a phenomenon poses to such research. A high-level overview of the experiment that utilises the two related methodologies is then presented in Section 6.4. The methodology to create company-specific corpora is then introduced

6.2 Colliding Cashtags

Cashtag collisions are incredibly common on Twitter, and often sow confusion for investors who are not aware of the issue (Evans, Owda, Crockett, and Ana Fernandez Vilas 2019). As of July 2021, 317 (15.8%) of the 2,006 companies on the LSE share a ticker symbol with companies listed just on the NASDAQ alone. A related issue that further adds to the confusion of this problem is the number of ways a ticker symbol can be used to refer to a stock-listed company online. Table 6.1 illustrates the different ways in which the PETS ticker can be used across different environments (Note: The ticker PETS refers to *Pets at Home Group PLC* on the LSE, and also refers to *PetMed Express Inc* on the NASDAQ). Exacerbating the issue further is the presence of cryptocurrencies - each of which possess their own ticker symbol which often collide with stock ticker symbols. There are currently over 6,000 cryptocurrencies in circulation according to CoinMarketCap[1], a website that actively monitors cryptocurrencies.

Time is a precious resource for investors, and the presence of cashtag collisions on Twitter only adds to the time it takes for investors to establish if the tweets returned by their cashtag searches are pertinent to the company in which they interested in. As shown in Figure 6.1, this phenomenon leads to investors often mistaking tweets that do not explicitly reference companies by name, but instead rely on the company ticker (cashtag) alone.

[1] https://coinmarketcap.com/

TABLE 6.1: The disparity of ticker symbols - Pets at Home Group PLC (LSE:PETS) and PetMed Express Inc (NASDAQ:PETS)

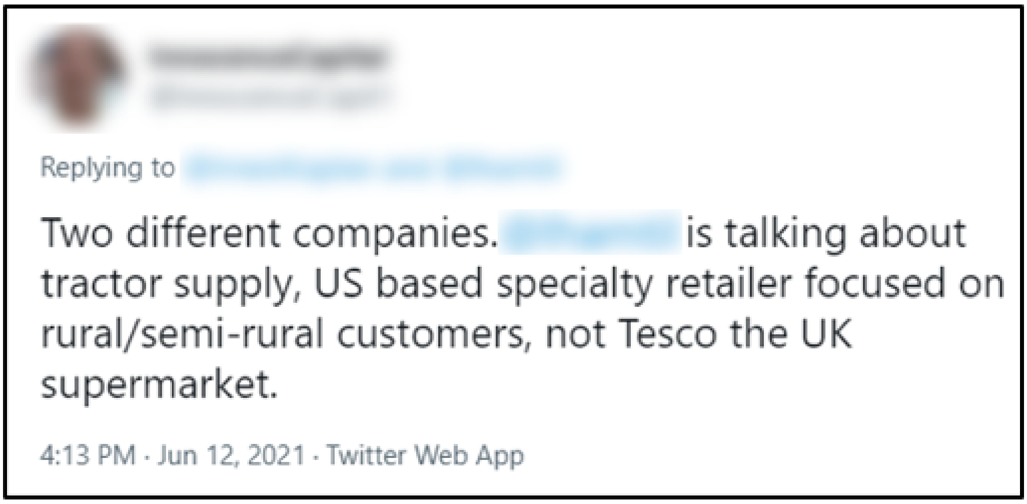
Exchange	**Reuters Intrument Code**	**Bloomberg Ticker**	**Google Finance Ticker**
LSE	PETSP.L	PETS:LN	LON: PETS
NASDAQ	PETS.O	PETS:US	NASDAQ: PETS

This research has identified two types of cashtag collisions:

1. At least two different companies listed on multiple exchanges use the same ticker symbol (and hence, the same cashtag). For example, $TSCO refers to Tesco PLC on the LSE, but also refers to the Tractor Supply Company listed on the NASDAQ.

2. The same company is listed on multiple exchanges, using the same ticker symbol. For example, $VOD is used to refer to Vodafone PLC on both the LSE and the NASDAQ.

Undoubtedly, the second type of cashtag collision will be harder to resolve, as companies with the same name will also feature many other similar elements, such as the terminology being used within tweets and the same leadership team (e.g. Chief Executive Officer). Resolving the first type of cashtag collisions will arguably be of more value - as it would allow both automated tools and investors to immediately ascertain that the tweet is irrelevant if it does not relate to the company they are searching for.



[redacted]

Replying to [redacted]

Two different companies. [redacted] is talking about tractor supply, US based specialty retailer focused on rural/semi-rural customers, not Tesco the UK supermarket.

4:13 PM · Jun 12, 2021 · Twitter Web App

FIGURE 6.1: Cashtag Collision Example

6.2.1 Challenges associated with Colliding Cashtags

Several challenges exist with attempting to use machine learning to resolve cashtag collisions. Most notable is the number of exchanges that may use a particular ticker symbol. The cashtag $WEB, for example, is a popular one given its name and is found on numerous exchanges around the world.

According to Cresci, Fabrizio Lillo, et al. (2018), automated spam bots are prevalent on Twitter. Cresci, Fabrizio Lillo, et al. (ibid.) collected over nine million tweets that contained at least one occurrence of a cashtag listed on one of the five major US financial markets over a five-month period. An interesting insight gleaned from this research is that users tweeting about low-value stocks would often include cashtags of high-value stocks - even if the tweet had no relevance to such high-value stock companies. The authors coined the term ″piggybacking″ to describe this behaviour, in which users would attempt to use the popularity of high-value cashtags to disseminate news about low-value stocks. Naturally, cashtags within a tweet that do not relate to the subject matter of the tweet could be considered noise, and could hinder efforts to resolve cashtag collisions. Their research concluded that almost 71% of retweets were made by automated accounts.

Another challenge associated with this work is applying Natural Language Processing (NLP) to tweets. Applying NLP to Twitter datasets is often challenging when compared to applying NLP techniques to structured documents (Alnajran and Technology 2019). Several such challenges outlined by Alnajran and Technology (ibid.) include:

- Tweets often contain acronyms and abbreviated forms of words in order to not exceed the 280-character limit
- A large amount of redundant information is circulated as people re-post (retweet) original messages
- Poor grammatical and syntactical structure, including misspelling are prevalent in micro-blogging messages
- Metadata within tweets (e.g. hashtags, cashtags, URLs) could interrupt the potential meaning of the tweet

The limited content in tweets (resulting from the character limitation) could be overcome by creating a custom corpus for each exchange-listed company that each contain terminology and keywords specific to that company. A company's corpus can then be consulted when training classifiers to predict is a tweet relates to a specific exchange-listed company or not.

6.3 Related work utilising Cashtags

As the issue of cashtag collisions has not been addressed within the literature, until our paper (Evans, Owda, Crockett, and Ana Fernandez Vilas 2019), this section will provide an overview of previous work which involve tweets containing cashtags. Previous works involving the analysis of cashtags could feature incorrect results and analyses due to the subtle nature of identifying and resolving cashtag collisions.

6.3.1 Disambiguation on Twitter

Resolving cashtag collisions can be seen as a disambiguation task, the aim of which is to attempt to remove any ambiguity as to what a tweet refers to. Several studies in the area of word disambiguation on Twitter exist (Gorrell, Petrak, and Bontcheva 2015; Inkpen et al. 2017; Spina, Gonzalo, and Amigó 2013). Spina, Gonzalo, and Amigó (2013) proposed a methodology to disambiguate company names within tweets on Twitter. The approach in Spina, Gonzalo, and Amigó (ibid.) involves associating positive and negative keywords with a company that, if found within the text of a tweet, assist in identifying which company is being referred to. For example, the word ″iPhone″ is considered a positive keyword for Apple on the NASDAQ, whereas the word ″pie″ would cause a negative shift in the tweet being associated with the Apple company. Positive and negative keywords were collected by scraping terms from company Wikipedia pages to build a corpus of keywords for each company to aid in the disambiguation task. Results from combining several weak models (bootstrapping) resulted in models obtaining up to 73% accuracy. One of the issues of relying on Wikipedia to generate such a corpus of keywords is that the language used within Wikipedia pages will not align with the informal 'slang' and abbreviations often found in tweets that are restricted to a character limit. In regards to the presence of cashtag collisions, this work may have unknowingly collected tweets relating to companies on other exchanges, potentially impacting the results obtained.

6.3.2 Stock Prediction using Cashtag Tweets

Rajesh and Gandy (2016) produced a system named CashTagNN, that uses sentiment and subjectivity scores of tweets that included the cashtags of two companies - Apple ($APPL) and Johnson and Johnson ($JNJ) on the NASDAQ - to model stock market movement. Tweets containing both of these cashtags were collecting and divided into two groups - tweets made whilst the market (NASDAQ) was open, and tweets made when the market was closed. A feed-forward neural network was then implemented that considered the sentiment scores for tweets within these categories to calculate the open and close market prices for these stocks. The authors reported a high accuracy when using sentiment values to predict the opening and close price of the stocks. A key issue was not addressed, however: If the AAPL and JNJ tickers are used on other exchanges, and the collection of tweets does not disregard tweets not relating to the NASDAQ companies, then this could make any findings susceptible to error.

6.4 Experiment Overview

Before the methodologies to resolve cashtag collisions (Sections 6.5 & 6.6) are introduced, a high-level overview of the experiment will be presented. This experiment (Figure 6.2) is aimed at resolving collisions for the 200 shortlisted companies that the SDE monitors discussion for (Appendix B).

The experiment consists of four phases: (1) experiment preparation, (2) data collection, (3) custom corpora creation created through data fusion, and (4) machine learning (classifier training). Each of these phases will now be summarised.

6.4.1 Phase 1: Experiment Preparation

The first step of the experiment involves selecting an exchange to resolve cashtag collisions for - naturally, this will be the exchange in which the research centres. For

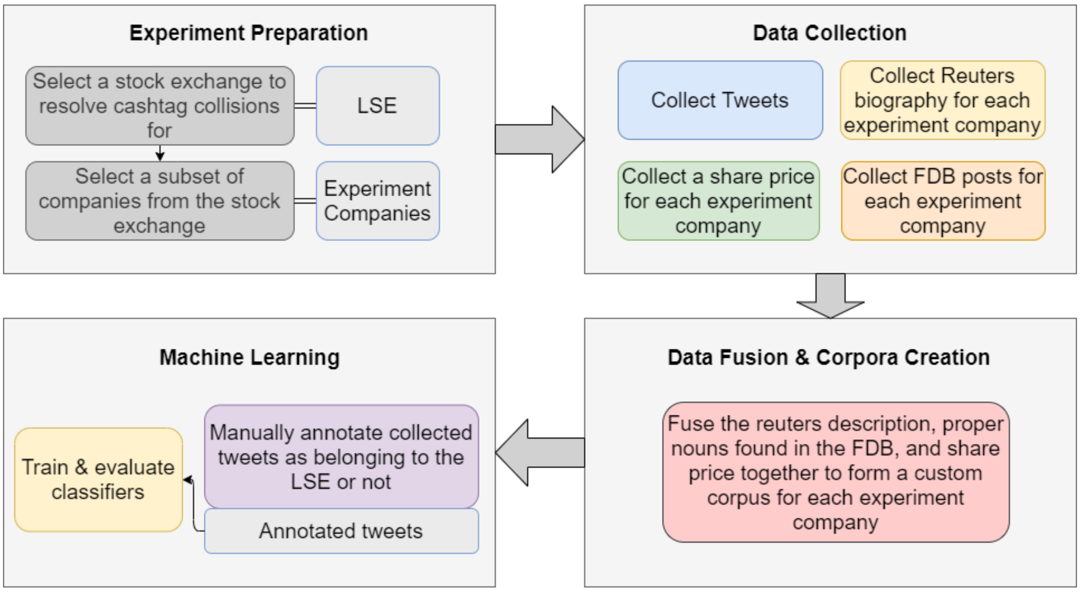

Experiment Preparation
Select a stock exchange to resolve cashtag collisions for
LSE
Select a subset of companies from the stock exchange
Experiment Companies
Data Collection
Collect Tweets
Collect Reuters biography for each experiment company
Collect a share price for each experiment company
Collect FDB posts for each experiment company
Machine Learning
Train & evaluate classifiers
Manually annotate collected tweets as belonging to the LSE or not
Annotated tweets
Data Fusion & Corpora Creation
Fuse the reuters description, proper nouns found in the FDB, and share price together to form a custom corpus for each experiment company

FIGURE 6.2: Cashtag collision experiment overview

the purposes of this research, the LSE is our chosen exchange. With the exchange selected, a subset of companies is then chosen (the companies previously shortlisted in Appendix B).

6.4.2 Phase 2: Data Collection

The second phase of the experiment involves the collection of tweets and other supplementary data to build the company-specific corpora. The data required for this experiment is summarised in Table 6.2, and expanded on below:

- **Tweets** - Tweets that contain at least one cashtag belonging to one of the experiment companies (Appendix B) will be collected using the Tweepy API (as discussed in Section 5.6.1)

- **Company Descriptions (via Reuters)** - The Reuters website contains a description of companies listed on all major stock exchanges around the world. Naturally, these descriptions will contain keywords that relate to the company (e.g. the company sector/industry, products and services they provide, countries of operation). Company descriptions for each of the experiment companies is collected via Scrapy, along with the CEO of the company (as this may aid the annotator if the CEO is mentioned within the tweet text).

- **London South East posts** - Discussion board posts will undoubtedly contain the terminology being used by investors when discussing LSE companies. We collect such posts from London South East for each of the shortlisted LSE companies in order to see what keywords are often used by investors. For example, based on research in Evans, Owda, Crockett, and Ana Fernandez Vilas (2019), it was found that Tesco PLC's (UK grocery company) discussion board featured mentions of Aldi, Lidl, and Sainsbury's - other grocery companies within the UK that are in direct competition with Tesco PLC. Naturally, the presence of such keywords in tweets could assist classifiers in determining if a tweet is indeed related to the LSE variant of the cashtag.
- **Share Prices (via AlphaVantage)** - As a subset of tweets will need to be annotated, a share price of the LSE company is collected and ultimately stored in a company-specific corpus (Section 6.5). This share price will assist in the annotation process, as some tweets may contain share prices which the annotator may be able to use to distinguish if the LSE company is being referred to. A recent share price is collected from AlphaVantage for each of the LSE experiment companies and stored within the corpora.

6.4.3 Phase 3: Data Fusion & Corpora Creation

Once the Reuters company descriptions, FDB posts, and share prices are collected, data fusion is utilised to build company-specific corpora for each of the LSE experiment companies. These corpora will contain keywords found on the LSE company Reuters description page, in addition to popular terminology used by investors on a popular financial discussion board (FDB) for the LSE company (Section 6.5).

6.4.4 Phase 4: Machine Learning

The machine learning phase first involves the manual annotation of tweets as belonging to the selected exchange (LSE) or not - a binary classification problem (Cryptocurrency tweets are labelled as not belonging to the LSE for the purposes of this experiment). These tweets were all labelled by myself by analysing the metadata associated with the tweet, including the contents (e.g. text, images and videos that

TABLE 6.2: Cashtag collision experiment data sources

Data Souce	Collected via	Features Collected	Date(s) Collected
Twitter	Tweepy	All metadata associated with the tweet	1/7/20 - 1/10/20 (3 months)
London South East Financial Discussion board posts	Scrapy	Post ID Subject Date Share Prices (at time of posting) Opinion Author Number of posts (of author) Premium member (true/false) Text	1/1/19 - 1/1/20 (1 year)
Reuters	Beautiful-Soup	Company Name Company Description Company CEO	1/10/20
AlphaVantage	Alpha-Vantage API	Share Price	1/10/20

may be embedded within the tweet), and details of the author (e.g. the username, user description, location of user). Tweets that did not contain enough information to be classified as belonging to the LSE (label: 1) were labelled as non-LSE (label: 0). Table 6.3 provides an example of a tweet from each of these labels, where the first tweet was labelled as being non-LSE (0), due to the only company being mentioned as belonging to the NASDAQ, with the other example tweet labelled as belonging to the LSE (1), due to the tweet containg a reference to a company that is only listed on the LSE. Traditional supervised machine learning classifiers are trained twice on each tweet (Section 6.6):

1. The first set of classifiers are trained solely on a bag of words (Section 6.6.2) of the tweet text, meaning each tweet is represented as a sparse vector.
2. The second set of classifiers are trained on the bag of words, in addition to features that are derived from the company corpus (Section 6.6.4).

The aim of these experiments were to classify whether a tweet was non-LSE (0) or LSE (1), and to obtain machine learning models that were capable of using selected

TABLE 6.3: Annotated tweet examples

Label	Tweet Class	Example Tweet Text
0	Non-LSE tweet	UBS Group Cuts Tractor Supply $TSCO Price Target to $97.00 https://t.co/TCGOolHQ9S
1	LSE tweet	$GGP - Greatland Gold PLC Exploration Update - Black Hills Drill Results https://t.co/Qk5wGpejIX"

features (discussed in Section 6.6.4).

6.5 Company Corpora Creation Methodology

This section will provide the methodology for creating company-specific corpora, and the natural language processing (NLP) techniques used on the data sources during this process. The purpose of this methodology is to build company-specific corpora for each of the LSE experiment companies in order to build a corpus of information (keywords and terms) specific to that company. The presence of such terms within the tweets' text will likely assist classifiers in determining if a tweet relates to the LSE or not, and the count of occurrences can effortlessly be converted to a feature to train such classifiers (6.6.5).

6.5.1 Corpora Creation

This section will detail the steps involved in creating company corpora (Fig 6.3).

Step 1: Feature Selection & Collection

The first step in creating the company-specific corpora is to select the features to be collected from each of the data sources (tweets, FDB Posts, and share prices), including the collection methods.

Step 2: Fusion Features

Although the Reuters company description and London South East posts contain several features which will be collected and stored, not all features available within these data sources will be advantageous to fuse.

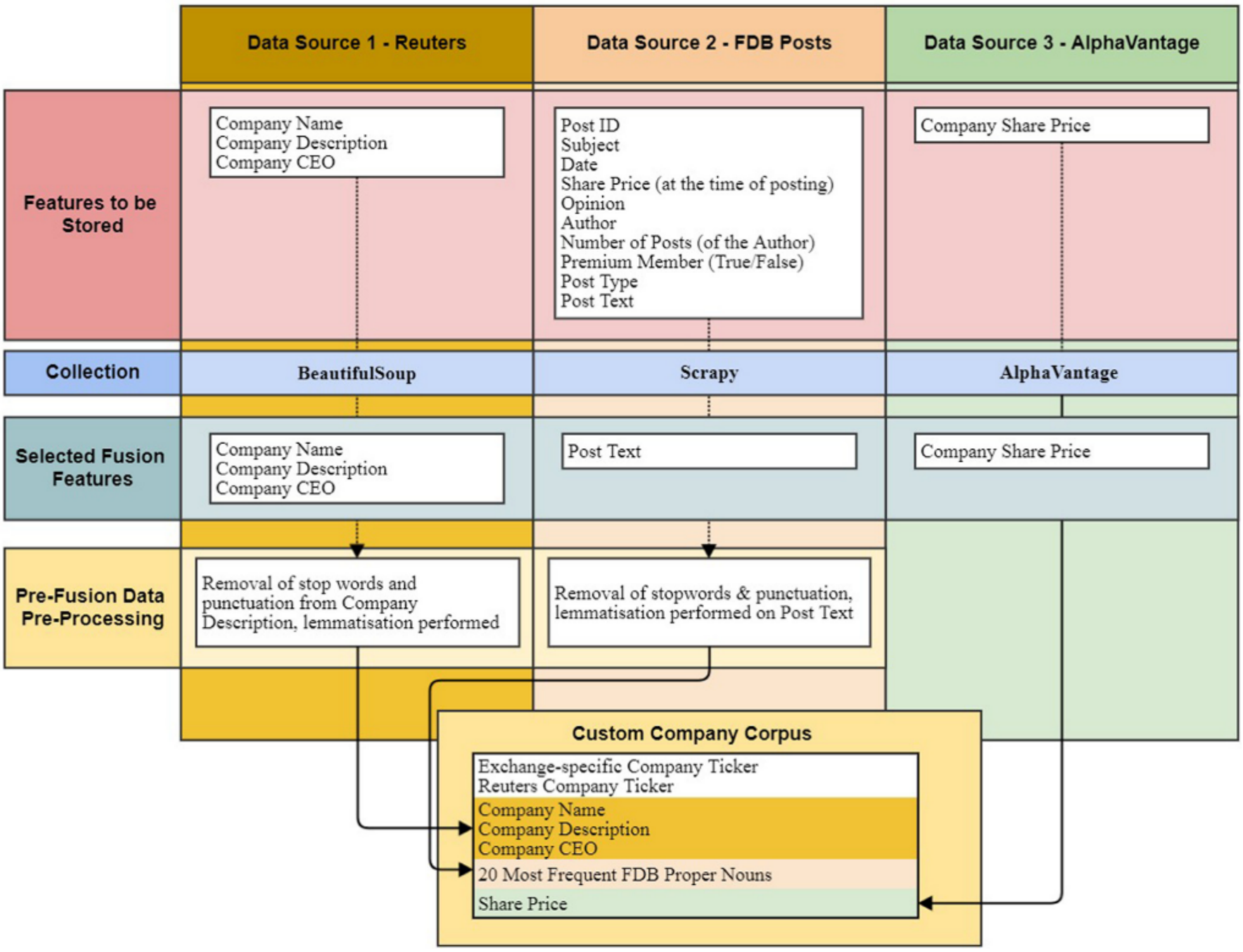

Data Source 1 - Reuters
Data Source 2 - FDB Posts
Data Source 3 - AlphaVantage
Features to be Stored
Company Name
Company Description
Company CEO
Post ID
Subject
Date
Share Price (at the time of posting)
Opinion
Author
Number of Posts (of the Author)
Premium Member (True/False)
Post Type
Post Text
Company Share Price
Collection
BeautifulSoup
Scrapy
AlphaVantage
Selected Fusion Features
Company Name
Company Description
Company CEO
Post Text
Company Share Price
Pre-Fusion Data Pre-Processing
Removal of stop words and punctuation from Company Description, lemmatisation performed
Removal of stopwords & punctuation, lemmatisation performed on Post Text
Custom Company Corpus
Exchange-specific Company Ticker
Reuters Company Ticker
Company Name
Company Description
Company CEO
20 Most Frequent FDB Proper Nouns
Share Price

FIGURE 6.3: Custom corpus creation methodology

Step 3: Data Pre-Processing

A crucial part of the fusion process is to remove redundant data that offers no benefit to being combined. The techniques described below have been utilised to meet this task. A summary of all of the pre-processing steps on each of the data sources is summarised in Table 6.4

Named Entity Recognition - Discussions taking place between investors on an FDB will contain countless words, some of which will be slang and casual discussion between the investors. To ensure that only relevant keywords are captured and stored within the corpus, we adopt Named Entity Recognition (NER) (Nadeau and Sekine 2007) to collect proper nouns found on the LSE company's London South East forum. Proper nouns are defined as ”a noun that designates a particular being or thing, does not take a limiting modifier, and is usually capitalized in English”[2].

[2] https://www.merriam-webster.com/dictionary/proper%20noun

In the context of FDB discussion, proper nouns will include the names of a company's competitors, names of relevant people to the company (e.g. CEO, celebrity endorsements), and the names of locations relevant to the company. NER is adopted to select the 20 most common proper nouns found within each company sub-forum on London South East. The count of such proper nouns in tweets can then be used as a feature when training classifiers to resolve cashtag collisions. For example, the Tesco corpus

Removal of stop words - Stop words are words that provide little or no value to a document (e.g. post), such as ”of”, ”the”, ”a”. Stop words have been identified and removed from all tweets, FDB posts, and Reuters company descriptions by Python's Natural Language Toolkit (NLTK)[3] package, which contains a comprehensive corpus of stop words.

Lemmatisation - The NLTK package has also been used to perform lemmatisaiton on the Reuters company descriptions, tweets, and FDB posts. Lemmatisation involves the grouping together of various inflected forms of a word into a single non-inflected word (Asghar et al. 2014). For example, the words ”playing”, ”plays”, and ”played” all have ”play” as their lemma. The primary purpose of lemmatisation in this experiment is to reduce the sparsity of the bag of words (discussed in Section 6.6.2).

TABLE 6.4: Data pre-processing techniques carried out on custom corpora data sources

	Data Source	Feature	Named Entity Recognition	Pre-processing Techniques		
				Stop word Removal	Lemmatisation	Other Removal
Fused Data Sources	Twitter	Tweet Text		✓	✓	Removal of URLs
	Financial Discussion Board Posts	Post Text	Proper Nouns (NNP)	✓	✓	
	Reuters	Company Description		✓		
	AlphaVantage	Share Price	No Pre-processing required			

6.6 Cashtag Collision Resolution Methodology (CCRM)

The methodology for resolving cashtag collisions (Figure 6.4) involves the creation of company-specific corpora, created through data fusion - the merging of different data sources together.

[3] www.nltk.org

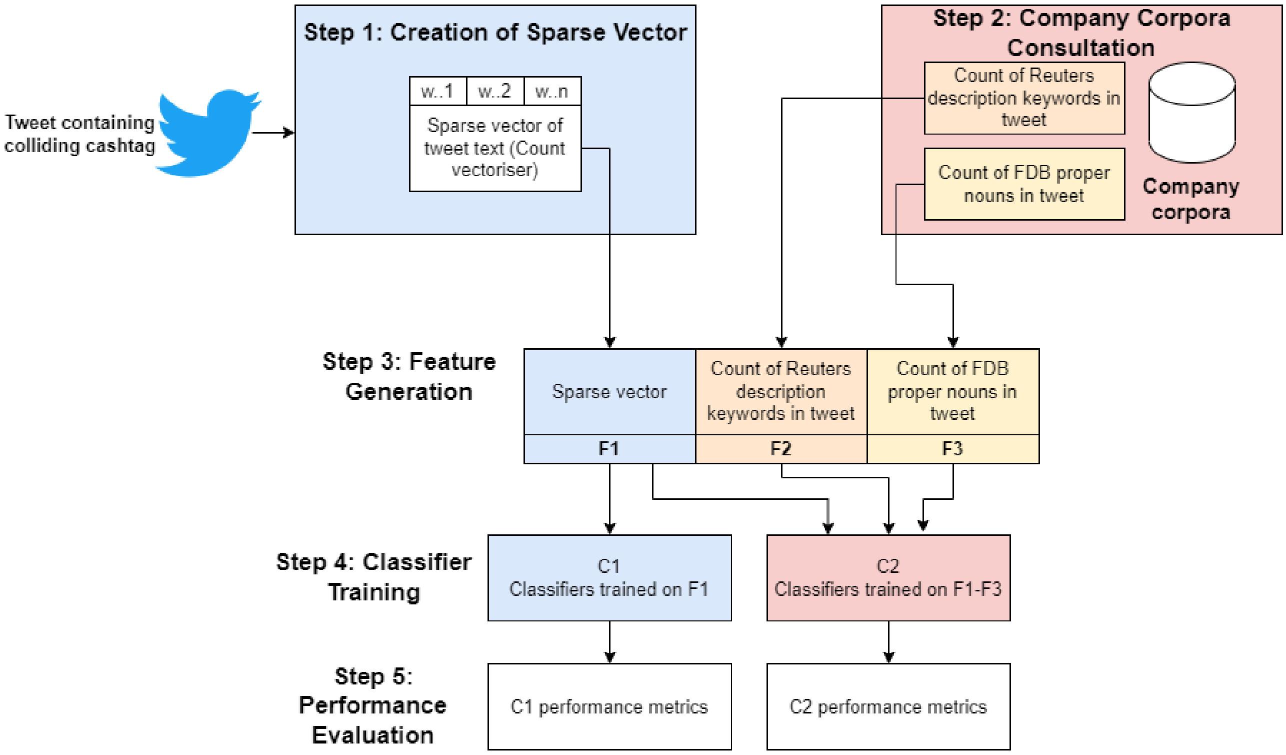

Step 1: Creation of Sparse Vector
Tweet containing colliding cashtag
w..1
w..2
w..n
Sparse vector of tweet text (Count vectoriser)
Step 2: Company Corpora Consultation
Count of Reuters description keywords in tweet
Count of FDB proper nouns in tweet
Company corpora
Step 3: Feature Generation
Sparse vector
Count of Reuters description keywords in tweet
Count of FDB proper nouns in tweet
F1
F2
F3
Step 4: Classifier Training
C1 Classifiers trained on F1
C2 Classifiers trained on F1-F3
Step 5: Performance Evaluation
C1 performance metrics
C2 performance metrics

FIGURE 6.4: Cashtag collision resolution methodology

6.6.1 Annotated tweet dataset

In total over the three-month window, 288,372 tweets were collected that contained at least one occurrence of a ticker symbol of an experiment company as defined in Chapter 4 and summarised in Appendix B. As previous research (Cresci, Fabrizio Lillo, et al. 2018) has found that spam bots and cryptocurrency tweets are widespread on Twitter, it is important that tweets are not randomly selected, as this could result in classifiers generalised to cryptocurrency tweets if such tweets are dominant within a dataset.

In our previous work (Evans, Owda, Crockett, and Ana Fernandez Vilas 2019), 1,000 tweets were shortlisted from a total of 86,539 tweets. The same approach to shortlist tweets is used for the SDE: we first attempt to select 25 tweets for each of the 200 SDE companies (for a potential total of 5,000 tweets). Using this selection criteria, 3,692 tweets were successfully shortlisted - meaning some company cashtags were not as popular as others. Tweets were then randomly selected from the pool of 288,372 to reach 5,000.

Tweets were annotated as belonging to one of two categories:

- **Non-LSE** - Tweets were labelled zero (0) if the tweet did not reference a company listed on the LSE. This include tweets relating to other exchanges, tweets containing cashtags but not referencing any stock, and tweets where there was not enough information to ascertain it was an LSE tweet.
- **LSE** - Tweets were labelled one (1) if the tweet contained some reference to the LSE company. For example, the name of the LSE company or GBP currency being referenced in the tweet. In the case of tweets that contain the 2nd type of cashtag collision (the same company listed on different exchanges), attributes such as currency and URLs within the tweet were considered.

In total, 3,120 of the 5,000 tweets were annotated as non-LSE (0), with the remainder (1,880 tweets) annotated as LSE tweets (1).

6.6.2 Step 1: Creation of tweet sparse vector

The first step of the CCRM (Figure 6.4) involves converting all of the annotated tweet texts (Section 6.6.1) into a sparse matrix (where each row is a sparse vector representing the text in each individual tweet), where w in Figure 6.4 is a single word in the sparse vector (and hence a feature in its own right), and n is the final word represented in the sparse vector.

6.6.3 Step 2: Company Corpora Consultation

As noted in Section 6.4.4, the first set of classifiers will be trained solely on the sparse vector (Section 6.6.2). The second set of classifiers, however, make use of features that are derived from a corpus specific to the company(s) in which the cashtag(s) of the tweet refer to. The second step, therefore, involves locating the relevant company corpora (based on the cashtag(s) within the tweet), in preparation for features to be generated, discussed next.

6.6.4 Step 3: Feature Generation

The features for each classifier set are then generated in preparation for training the classifiers.

1. First set of Classifiers (C1) - trained solely on the sparse vector of the tweet text (Feature 1 - F1). In order to reduce the sparsity of this sparse matrix, stop words are removed from the tweets, and lemmatisation is performed to bring words to their non-inflected forms.

2. Second set of Classifiers (C2) - trained on F1 (the sole feature of C1), and the count of Reuters description keywords in tweet (F2), and the count of FDB proper nouns (Section 6.5.1) in tweet (F3). For example, if a tweet contains Tesco's cashtag ($TSCO) and contains the words Sainsburys and Aldi once each respectively, then the count for F3 would be two for that specific observation (tweet).

6.6.5 Step 4: Classifier Training

The two groups of classifiers are then trained (Section 6.7.2) on the two different feature sets. Traditional supervised machine learning classifiers that have enjoyed success in various tweet classification tasks (Verma and Sofat 2014; Evans, Owda, Crockett, and Ana Fernandez Vilas 2019) are employed, these include: Naive Bayes, Logistic Regression, Decision Tree, Random Forest, k-Nearest Neighbours, Support Vector Machine. The reason for relying on such traditional classifiers, and not on deep learning, is that deep learning relies on larger datasets - which can be difficult to obtain for supervised learning due to the cost and time associated with supervised learning. The results of those classifiers are then evaluated and discussed in the next step.

6.6.6 Step 5: Performance Evaluation

The final step of the CCRM (Figure 6.4) includes comparing each of the classifiers to ascertain if the additional features derived from the company corpora provide additional informative power to the classifiers to correctly predict if a tweet belongs to the experiment exchange (LSE) or not. As the annotated dataset is imbalanced, Section 6.7.1 will detail how these classifiers are evaluated using a metric ideal for binary classifiers trained on an imbalanced dataset.

6.7 Results & Discussion

This section will present the results of both experiments (where each set of classifiers map to an experiment) to determine if the inclusion of features derived from the custom company corpora yield any benefit to the classifiers.

6.7.1 Classifier Evaluation

Before the results are presented, it is important to mention why the accuracy of the classifiers is not used to judge the performance of the classifiers. As the dataset used to train each group of classifiers is imbalanced, accuracy can give a false indication of classifier performance. In the case of the 5,000 tweets used in this experiment, 3,120 are labelled as not relating to the LSE. This means that if we were to abandon the machine learning model approach, and simply predict zero every time, an accuracy of 62% would be achieved. This is referred to as the accuracy paradox and is particularly problematic where certain classes are incredibly rare and hence not well-represented within the dataset (Valverde-albacete and Pela 2014).

Several solutions exist to the accuracy paradox, these include:

- **Balance dataset using over/under-sampling techniques** - The dataset could be balanced so that each class is equally represented. Techniques to achieve this include over-sampling (e.g. SMOTE) (Chawla et al. 2002) and under-sampling (e.g. ACOS) (Yu, Ni, and Zhao 2013), whereby the former creates new samples of the minority class, and the latter involves reducing the number within the majority class (Rong, Gong, and X. Gao 2019).

- **Manually balance dataset** - Additional annotation could be undertaken to bring the minority class (1 - LSE tweets) to the same sample size as the majority class (0 - non-LSE tweets). Such an approach would naturally lengthen the time of the experiment considerably due to the laborious nature of manual annotation (Hsueh, Melville, and Sindhwani 2009).

- **Evaluate classifiers using metrics that account for class imbalances** - The precision and recall of the classifiers, including the f1-score (harmonic mean of precision and recall) could be used to compare each of the classifiers.

For the purposes of these experiments, a metric that is especially suited for binary classifiers trained on an imbalanced dataset was chosen to judge the performance of the classifiers: the Matthews Correlation Coefficient (MCC) score (Chicco and Jurman 2020). The MCC score is calculated directly from the classifier confusion matrix (CM), using Equation 6.1, where:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6.1}$$

- **TP** = True Positive: An LSE tweet was successfully predicted as relating to the LSE.
- **TN** = True Negative: A non-LSE tweet was successfully predicted as not relating to the LSE
- **FP** = False Positive: A non-LSE tweet was incorrectly predicted as relating to the LSE. Also known as a Type I error.
- **FN** = False Negative: An LSE tweet was incorrectly predicted as being a non-LSE tweet. Also known as a Type II error.

A value of -1 to +1 is returned as a result of applying the equation to the confusion matrix values. An MCC score of -1 indicates the classifier has made incorrect predictions on all observations, with a score of +1 indicating the classifier has made correct predictions across all observations (Liu et al. 2015).

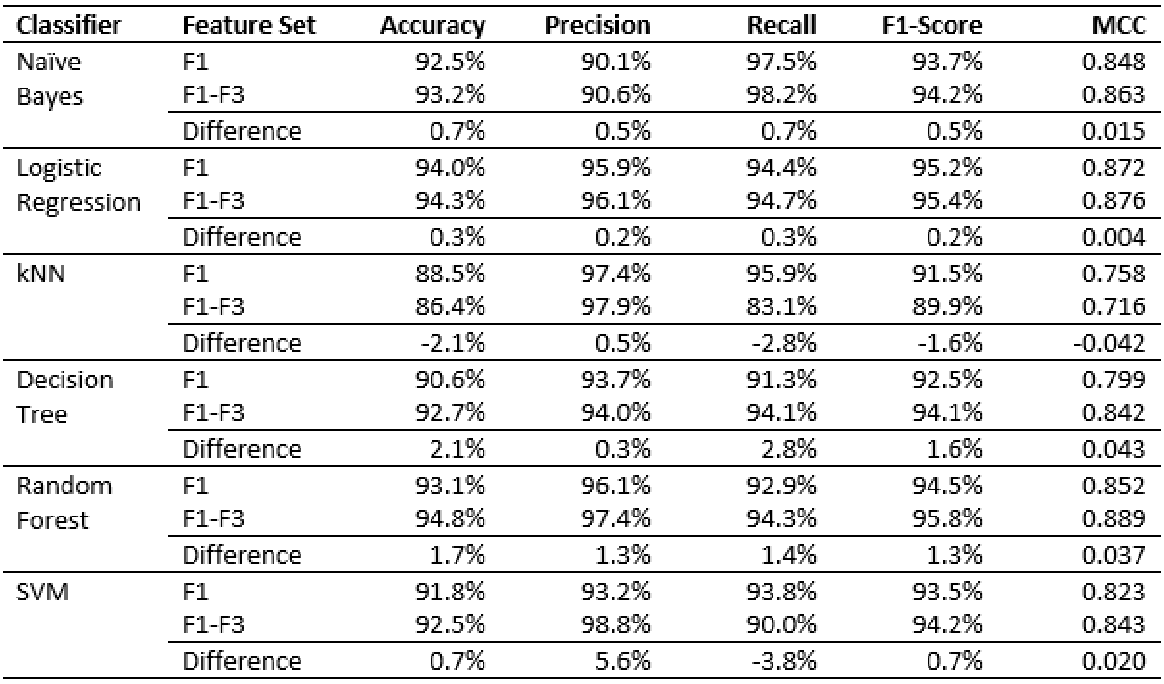
6.7.2 Classifier Results

The machine learning classifiers trained in these include: Naive Bayes (NB), Logistic Regression (LR), k-Nearest Neighbours (kNN), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF).

All classifiers were implemented using the scikit-learn[4] library in Python, using an 80/20 train/test split and 10-fold cross-validation. Optimal hyperparameters for

[4] https://scikit-learn.org/stable/index.html

TABLE 6.5: Cashtag collision classifier results



Classifier	Feature Set	Accuracy	Precision	Recall	F1-Score	MCC
Naïve Bayes	F1	92.5%	90.1%	97.5%	93.7%	0.848
	F1-F3	93.2%	90.6%	98.2%	94.2%	0.863
	Difference	0.7%	0.5%	0.7%	0.5%	0.015
Logistic Regression	F1	94.0%	95.9%	94.4%	95.2%	0.872
	F1-F3	94.3%	96.1%	94.7%	95.4%	0.876
	Difference	0.3%	0.2%	0.3%	0.2%	0.004
kNN	F1	88.5%	97.4%	95.9%	91.5%	0.758
	F1-F3	86.4%	97.9%	83.1%	89.9%	0.716
	Difference	-2.1%	0.5%	-2.8%	-1.6%	-0.042
Decision Tree	F1	90.6%	93.7%	91.3%	92.5%	0.799
	F1-F3	92.7%	94.0%	94.1%	94.1%	0.842
	Difference	2.1%	0.3%	2.8%	1.6%	0.043
Random Forest	F1	93.1%	96.1%	92.9%	94.5%	0.852
	F1-F3	94.8%	97.4%	94.3%	95.8%	0.889
	Difference	1.7%	1.3%	1.4%	1.3%	0.037
SVM	F1	91.8%	93.2%	93.8%	93.5%	0.823
	F1-F3	92.5%	98.8%	90.0%	94.2%	0.843
	Difference	0.7%	5.6%	-3.8%	0.7%	0.020

each classifier were obtained by undertaking a grid search. The results of the experiments on the different feature sets are presented in Table 6.5. Each of the classifiers will now be discussed in turn, including a discussion on the top-performing classifiers and which classifier is deployed in the SDE to resolve cashtag collisions.

Naive Bayes

The first classifier trained was a Multinomial NB classifier due to its suitability with text classification tasks (S.-B. Kim et al. 2006). The results (Table 6.6 - which shows the confusion matrix table and MCC score of both classifiers) show a marginal improvement when the NB classifier is trained on the combined feature groups.

TABLE 6.6: Naive Bayes results

	Sparse Vector (F1)		**Combined Features (F1-F3)**	
CM	559	61	562	58
	14	366	10	370
MCC Score	0.848		0.863	

6.7.3 Logistic Regression

A LR classifier was then trained on both feature sets, with the results contained in Table 6.7. LR are particularly suitable for tasks with a dichotomous outcome (Mood 2010) (in this case, a tweet referring to the LSE or not). The results of training LR classifiers on the different feature sets indicate that the inclusion of features derived from the company corpora do not provide a significant performance increase.

TABLE 6.7: Logistic Regression results

	Sparse Vector (F1)	**Combined Features (F1-F3)**
CM	595 25	596 24
	35 345	33 347
MCC Score	0.872	0.876

6.7.4 k-Nearest Neighbours

A kNN classifer was then trained on both features sets (Table 6.8). Out of all of the classifiers trained in this experiment, the kNN classifier was the only classifier to be negatively affected when additional features derived from the company corpora were included.

TABLE 6.8: kNN Results

	Sparse Vector (F1)	**Combined Features (F1-F3)**
CM	604 16	607 13
	99 281	123 257
MCC Score	0.758	0.716

6.7.5 Decision Tree

Next, a DT classifier was trained on the feature sets, with the results reported in Table 6.9. DTs are considered one of the major success stories within the AI community due to their ease of interpretation and ability to be visualised (Freund and Mason 1999; Vadera 2010). The results indicate that the DT trained on the combined feature groups yield a higher performance than being trained on the sparse vector of the tweet text alone.

TABLE 6.9: DT Results

	Sparse Vector (F1)	**Combined Features (F1-F3)**
CM	581 39	583 37
	55 325	36 344
MCC Score	0.799	0.842

6.7.6 Random Forest

Table 6.10 presents the results from training a RF classifier on each of the feature sets. This classifier in particular has seen a significant rise in the MCC score as a result of being trained on the combined features of the BoW and the corpora features, with less Type I and Type II errors reported.

TABLE 6.10: RF Results

	Sparse Vector (F1)	**Combined Features (F1-F3)**
CM	596 24	604 16
	45 335	36 344
MCC Score	0.852	0.889

6.7.7 Support Vector Machine

Lastly, a SVM classifier was trained on each of the feature sets (Table 6.11). Although the SVM on the combined feature set does not yield an increase as significant of that of the RF or several of the other classifiers, it is still yields a slight improvement in the MCC score. Interestingly the number of Type I errors (false positives) is reduced substantially (meaning less non-LSE tweets are incorrectly predicted as relating to the LSE), whereas the number of Type II errors (false negatives) significantly rises (more LSE tweets are incorrectly predicted as being LSE).

TABLE 6.11: SVM Results

	Sparse Vector (F1)	**Combined Features (F1-F3)**
CM	590 43	613 7
	39 328	68 312
MCC Score	0.823	0.843

6.7.8 Discussion

Based on the results in Table 6.5, almost all of the classifiers, with kNN being the exception, saw an improvement in their MCC score when features derived from the company corpora was included in the feature set. The two best-performing classifiers in respect to their MCC score were LR (0.876) and RF (0.889). The RF and DT classifiers also benefit from being easier to interpret, as the decisions the algorithms take can be visualised by producing a visual output of the tree with its various nodes and decisions. Tree-based model are also more robust to overfitting and less computationally expensive to train than newly designed approaches in the literature such as SVM (Parmezan, H. D. Lee, and Wu 2017). As the RF classifier possesses the highest MCC score, and therefore is able to resolve a higher proportion of cashtag collisions, this classifier has been deployed in the third level of the data fusion model to ensure irrelevant tweets are not carried forward.

6.8 Chapter Summary

This chapter has explored a critical task of the data fusion model: resolving cashtag collisions present in tweets. It is important to highlight that if such collisions are not addressed, then any attempt to identify irregular tweets (e.g. clustering, to be discussed in Section 8.6) will involve clustering of different categories (e.g. LSE, non-LSE, and cryptocurrency tweets), instead of tweets that are like-for-like (LSE tweets). With the issue of cashtag collisions now addressed, the next chapter will introduce the detection capabilities of the ecosystem, which make use of clustering algorithms to detect irregular posting activity surrounding financial stock discussion.

The main contributions of this chapter include:

- A novel methodology to create company-specific corpora, in which features can be derived to assist classifiers to resolve cashtag collisions.
- A novel methodology for resolving tweets containing colliding cashtags that utilises machine learning classifiers trained on the tweet text, and features derived from company-specific corpora.

- Evidence that the inclusion of features derived from company corpora lead to better-performing classifiers in respect to metrics that account for imbalanced class distributions (MCC score).

The next chapter will detail the final contribution to the data layer: assessing the credibility of financial stock tweets.

Chapter 7

Data Layer: Assessing Tweet Credibility

7.1 Overview

This chapter will provide an overview of how the Smart Data Ecosystem (SDE) (introduced in Section 4.2) assesses the credibility of financial stock tweets. After tweets are collected by the SDE, and filtered (non-LSE tweets and cryptocurrency tweets are discarded), the credibility of such financial tweets are then assessed, the focus of this chapter. The aim of this chapter is to address the first research question posed in Section 1.5: *Can a smart data ecosystem, utilising machine learning classifiers, classify social media posts with respect to their credibility?*

Investments are often made as a result of timely and credible information being made available to investors. Since Twitter's inception of the cashtag feature in 2012, it has seen increased use by investors to discuss and disseminate news surrounding stocks (Ranco et al. 2015). The term *credibility* is generally defined as "the believability of information" (Sikdar et al. 2013), with social media credibility defined as "the aspect of information credibility that can be assessed using only the information available on a social media platform" (C. Castillo, Mendoza, and Poblete 2011). Assessing the credibility of financial stock tweets is particularly challenging due to exchanges and regulators need to quickly curb the spread of misinformation that may be circulating online surrounding stocks. Specifically, Twitter users who attempt to capitalise on the fast dissemination that Twitter provides may become susceptible to apocryphal information that is circulating on such a platform, further highlighting the need for mechanisms to assess the credibility of messages. Twitter does not only act as a discussion platform for investors, but also as an aggregator for financial information by companies and regulators. The financial market community is currently bereft of ways of assessing the credibility of financial stock tweets, as previous work on credibility within Twitter has focused on areas such as politics and natural disaster events (Alrubaian et al. 2018). This chapter presents research to bridge that gap - supervised classifiers are trained (Section 7.6) on a novel set of general and financial features (Section 7.4) to assess the credibility of financial stock tweets.

Firstly, the related work on credibility is introduced (Section 7.2). The methodology utilised by the SDE for assessing the credibility of financial stock tweets is then provided (Section 7.3). An overview of the different feature groups considered by the different classifiers is then given (Section 7.4). The feature selection techniques utilised as part of the classifier training process is outlined in Section 7.5. An experiment designed to validate the methodology is then presented (Section 7.6).

7.2 Related work on Credibility

Existing research on assessing the credibility of financial stock tweets is scant within the literature, as much of the existing research on credibility on Twitter is geared towards areas such as natural disaster events (J. Yang et al. 2019), healthcare (Bhattacharya et al. 2012), and politics (Sikdar et al. 2013; Page and Duffy 2018). Alrubaian et al. (2018) undertook an extensive survey of previous work on assessing the credibility of microblogging messages, in which they looked at 112 papers on the subject over the period of 2006-2017. One of the key changes cited by Alrubaian et al. (ibid.) is that there is a large amount of literature that has developed different credibility dimensions and definitions, meaning a unified definition of what constitutes credible information does not exist. The majority of previous work on credibility is based on supervised approaches, such as Support Vector Machines (SVM) and Bayesian algorithms, which will now be explored (ibid.).

The pioneering work on assessing tweet credibility is attributed to C. Castillo, Mendoza, and Poblete (2011), in which they assessed the credibility of tweets during a two month window relating to current news events. Their research demonstrated the success of using classifiers such as Naive Bayes, Support Vector Machine, and Logistic Regression to classify tweets as falling into one of four classes: (1) almost certainly not true, (2) likely to be false, (3) almost certainly true, and (4) undecided. Up to 89% of topic appearances and their associated credibility classification achieved precision and recall scores of up to 80%.

Much of the research undertaken since the work of (ibid.) has built upon their successes of using machine learning to classify microblogging posts' credibility. Morris et al. (2012) carried out a series of experiments to ascertain which features provided the most informative powers to classifiers when assigning credibility to tweets. Many of the features Morris et al. (ibid.) found to be particularly useful for assessing credibility were primarily user-based features (e.g. user's reputation as indicated by their verified status, and the user's description). Morris et al. (ibid.) conducted a follow-up experiment in which they found that the topic of a message affected the perception of credibility, with tweets relating to science found to be more credible. Another insight from the research of Morris et al. (ibid.) is the impact a user's profile picture on assessing credibility, with Twitter users who have the default Twitter profile image (assigned when the account is created) perceived to be less credible than users who have changed their profile image.

User-based features (e.g. the number of followers a user has) have been examined intently within the literature as a means of assessing credibility (Alrubaian et al. 2018). Depending solely on such features, however, has faced criticism, as Twitter users are able to buy followers, leading to an artificial increase in their follower base, and therefore leading to a false impression of credibility (De Micheli and Stroppa 2013; Cresci, Di Pietro, et al. 2015).

Hassan et al. (2018) developed a credibility detection model that was based on machine learning techniques and employed an annotated dataset of news events annotated by a team of journalists. Two feature groups were developed (1) features derived from the content (e.g. length of the tweet text), and (2) features derived from the source (e.g. does the user still possess the default profile picture?). Three groups of classifiers were trained: (1) classifiers trained on the content feature group, (2) classifiers trained on the source feature group, and (3) classifiers trained on both feature groups. The researchers demonstrated that the classifiers trained on the third feature group (content and source features), performed better than individual feature groups alone. However, the researchers neglected to test if the performance of the two best-performing classifiers was statistically significant.

As the topic of credibility is a subjective one, researchers have tried to assess the impact of bias when annotating a subjective annotation task. Bountouridis et al. (2019) considered the bias involved relating to dataset annotation around the area of credibility. The researchers found that biases are particularly prevalent in annotated credibility datasets. Factors such as population, external, cultural, and enrichment biases all impact an annotator′s decision making process, and hence their annotation choices. As with other subjective tasks, the data is annotated by specific people, with a specific worldview, at a specific time, making specific methodological choices (ibid.). When an annotation task is subjective, studies have often depended on the ′wisdom of the crowd′, whereby multiple annotations are sought by different individuals, and the majority opinion is used to reach a consensus (Sikdar et al. 2013; C. Castillo, Mendoza, and Poblete 2011; El Ballouli et al. 2017; Lorek et al. 2015). In cases where a majority cannot be determined, observations could be removed or given to a final decision maker who makes the final annotation judgement (Sikdar et al. 2013; Gupta and Kumaraguru 2012).

The use of crowdsourcing platforms have proved popular over the years as a means of leveraging the opinion of a large number of annotators. Platforms such as Amazon′s Mechanical Turk[1], and Appen[2] (formerly Figure Eight) provide services in which annotations can be obtained from their vast network of members. The use of such crowdsourcing services has faced some criticism in recent years, as the the annotators on such platforms often do not possess the expected domain knowledge for the specific annotation tasks (ODonovan et al. 2012; M.-C. Yang and Rim 2014). Alrubaian et al. (2018) argue that depending on the wisdom of the crowd in this way is not ideal, as the lack of domain knowledge could lead to obtaining bad-quality annotations.

[1] https://www.mturk.com/

[2] https://appen.com/

Much of the work undertaken on assessing credibility has been performed offline in a post-hoc setting, whereby tweets are collected, annotated, and then used to train classifiers. Gupta, Kumaraguru, et al. (2014) designed and developed a plug-in for the Google Chrome browser capable of assigning credibility scores to tweets as they are published to the platform. The score ranged from 1 (lowest) to 7 (highest), and was produced by a semi-supervised algorithm trained on human labels obtained through crowdsourcing and considered over 45 features. The plug-in was evaluated in terms of response time (time taken to retrieve the credibility score for a tweet), usability, and effectiveness were evaluated on a dataset of 5.4 million tweets. The results of the system evaluation demonstrated that 63% of users surveyed either agreed with the credibility score, or disagreed by 1-2 points.

7.2.1 Summary of Related Work

The issue of much of the related work on assessing tweet credibility lies in the fact that researchers do not provide the predictive power of features used in the training of classifiers. Naturally, classifiers that are particularly susceptible to the curse of dimensionality (e.g. k-Nearest Neighbours), suffer decreased performance as more features are considered (Parmezan, H. D. Lee, and Wu 2017). As a result, many of the features proposed for assessing credibility could be irrelevant (particularly if such features are not omnipresent in financial stock tweets), which could lead to models overfitting. The explored works in this chapter typically group features into different categories (e.g. tweet/content features, and user/author features), and the credibility classification is assigned to a tweet, or the author of the tweet. As mentioned previously, user features, such as a user's number of followers, can be artificially inflated, giving a false indication of credibility, meaning taking into consideration other features relating to the tweet and the author is important.

The methodology to be adopted by the SDE (Section 4.2) for assessing the credibility of stock tweets (Section 7.3) will highlight and discard irrelevant features during the training of the classifiers to alleviate such concerns, and report which features are particularly informative for assessing credibility.

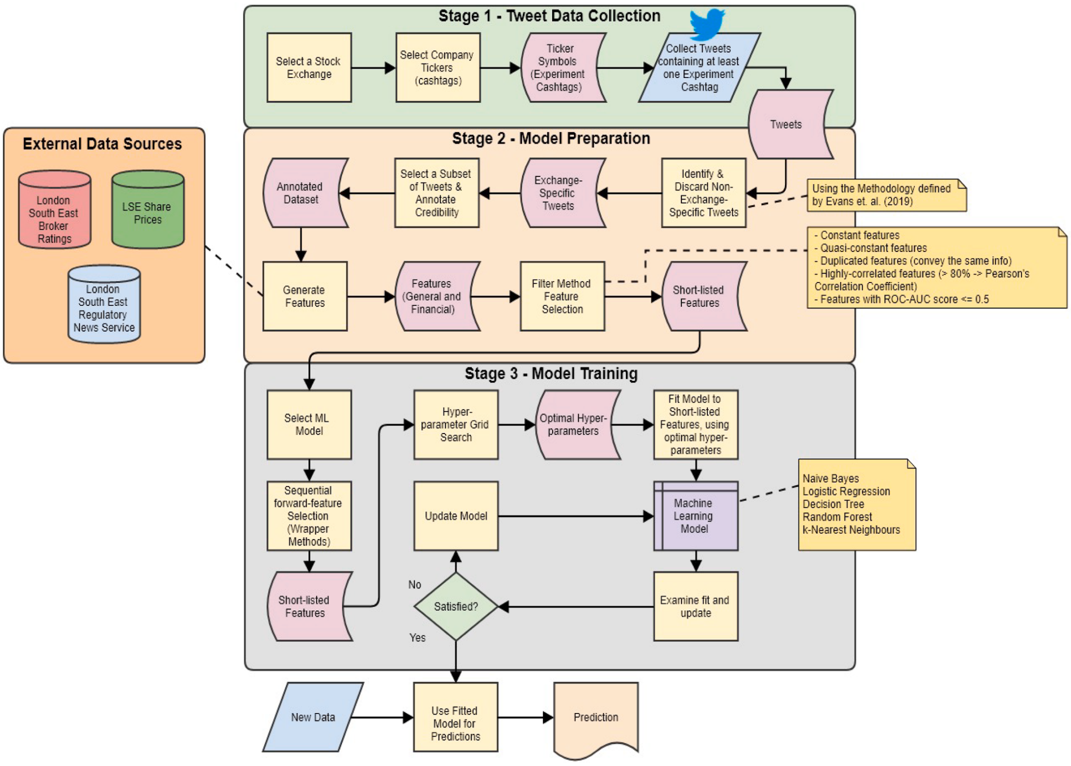

Stage 1 - Tweet Data Collection
Select a Stock Exchange
Select Company Tickers (cashtags)
Ticker Symbols (Experiment Cashtags)
Collect Tweets containing at least one Experiment Cashtag
Tweets
External Data Sources
London South East Broker Ratings
LSE Share Prices
London South East Regulatory News Service
Stage 2 - Model Preparation
Annotated Dataset
Select a Subset of Tweets & Annotate Credibility
Exchange-Specific Tweets
Identify & Discard Non-Exchange-Specific Tweets
Using the Methodology defined by Evans et. al. (2019)
Generate Features
Features (General and Financial)
Filter Method Feature Selection
Short-listed Features
- Constant features
- Quasi-constant features
- Duplicated features (convey the same info)
- Highly-correlated features (> 80% -> Pearson's Correlation Coefficient)
- Features with ROC-AUC score <= 0.5
Stage 3 - Model Training
Select ML Model
Hyper-parameter Grid Search
Optimal Hyper-parameters
Fit Model to Short-listed Features, using optimal hyper-parameters
Sequential forward-feature Selection (Wrapper Methods)
Update Model
Machine Learning Model
Naive Bayes
Logistic Regression
Decision Tree
Random Forest
k-Nearest Neighbours
Short-listed Features
No
Satisfied?
Yes
Examine fit and update
New Data
Use Fitted Model for Predictions
Prediction

FIGURE 7.1: Credibility Assessment Methodology (CAM)

7.3 Credibility Assessment Methodology

This section presents the Credibility Assessment Methodology (CAM) for assessing the credibility of financial stock tweets. The CAM (Figure 7.1) consists of three phases: (1) data collection, (2) model preparation, and (3) model training. The following subsections will provide a high-level overview of this methodology, with specific implementation details discussed in Section 7.6, where the CAM is adopted to assess the credibility of tweets pertaining the the LSE.

7.3.1 Stage 1 - Data Collection

The first stage of the CAM is the selection of a stock exchange and the shortlisting of companies for that exchange in which to assess the credibility of tweets for. Once an exchange and a list of companies has been selected, the collection of tweets can commence using a suitable API.

7.3.2 Stage 2 - Model Preparation

The second stage of the methodology is concerned with disregarding irrelevant tweets, selecting and generating features, and highlighting features that do not offer much - or any - predictive power to classifiers.

Tweet Filtering

Firstly, the model preparation stage must identify and discard collected tweets that do not correspond to the selected stock exchange. This is achieved using the cashtag collision resolution methodology discussed in Section 6.6.

Tweet Annotation

As supervised machine learning models are to be trained to assess the credibility of stock tweets, an annotated dataset must be created. As discussed in the related work (Section 7.2), researchers treat this as either a binary classification problem (i.e the tweet is either credible or not), or include more labels for more granularity. Section 7.6.3 provides a detailed overview of how the annotation process was undertaken within the experiment, along with a justification of the annotation procedure.

Feature Engineering & Selection

Once a dataset of tweets has been annotated to the pre-determined credibility classes, features can be engineered and selected in preparation for the classifier training process. Filter-based feature selection techniques are employed to identify features that offer little or no informative power to the credibility classifiers, in an attempt to reduce the feature space to create more robust classifiers (Rong, Gong, and X. Gao 2019). Such features may include those that are constant (the same across all observations), quasi-constant (the same across almost all observations), or duplicated features that convey the same information (Bommert et al. 2020). A full description of the feature selection techniques employed in this methodology are reserved for Section 7.5.

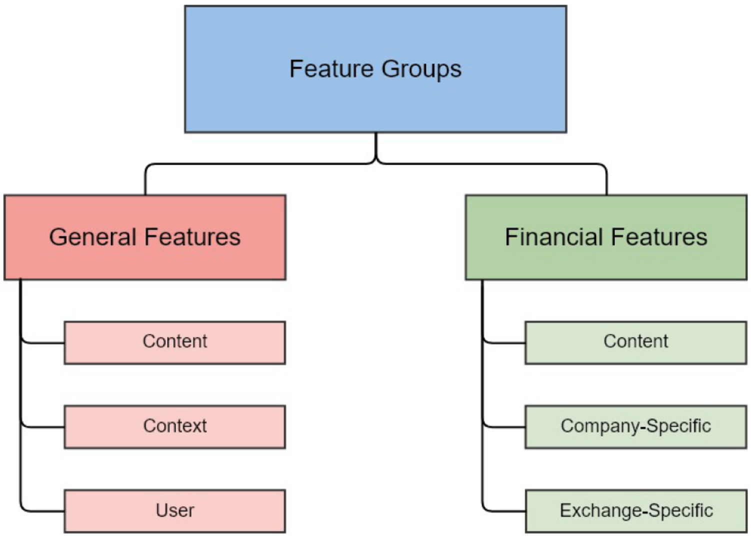


FIGURE 7.2: Credibility feature groups

7.3.3 Stage 3 - Model Training

The last stage of the methodology involves conducting additional feature selection techniques through repeated training of classifiers to identify optimal feature sets by adopting wrapper selection feature selection techniques (Wah et al. 2018). Once an optimal feature set has been identified for each classifier, hyperparameter grid searches are conducted on the classifiers that have tunable hyperparameters (all except Naive Bayes) in order to find further performance increases.

7.4 Feature Groups

This section presents the two feature groups that are used to train the classifiers. The features proposed are divided into general features (GFs) and financial features (FFs). The full list of features considered can be found in Appendix F. It is anticipated that not every feature will offer an equal amount of informative power to the classifiers to be trained, meaning we do not attempt to justify each of the features, but instead remove features that are found to be of little or no benefit to the classifiers. The general and financial feature groups, including their corresponding sub-groups are depicted in Fig 7.2.

7.4.1 General Features

General features play an important part in assessing tweet credibility. Such features can be created from any tweet (financial or otherwise). This research divides GFs into three sub-groups: (1) content features, (2) context features, and (3) user features. Each of these feature sub-groups will now be discussed further.

Content

Content-based features are those that can be either directly derived from the tweet text, or engineered from the text in some way. Features in this group include the count of different keyword groups present in the tweet text (e.g. noun, verb), including details of hyperlinks found within the tweet text (e.g. does the tweet contain a reference to a popular website). The motivation for this group relates to the second dimension of tweet credibility - the credibility of information within the tweet.

Context

Context-based features include information about the tweet that is not relating to the content or user, but focuses on information such as *when* the tweet was published to twitter. Naturally, the mere presence of a hyperlink within the tweet should not be a sign of the tweet being credible, as the hyperlink may be completely irrelevant to the tweet, or may be a dead hyperlink (does not navigate to a live page). One of the features in this group includes the ″count of live URLs within the tweet″, which involves visiting each hyperlink within the annotated dataset of tweets. A live URL is defined as any URL that returns a successful response code (2XX). Another feature is the number of popular URLs contained within the tweet, as determined by moz[3], a website that ranks the popularity of domains. There are several ways of publishing a message to twitter, these typically fall under the categories of manual and automatic. Manual methods include a user typing their tweet and manually publishing via a mobile device or computer. Automatic methods, on the other hand, involve the publishing of tweets based on rules and triggers (e.g. specific time of the

[3] https://moz.com/top500

day/week) (S. Castillo et al. 2019). Popular frameworks and providers for providing automatic tweet publishing include TweetDeck[4], Hootsuite[5], and IFTTT[6].

User

User-based features have enjoyed considerable success in the literature as a means of assessing the credibility of tweets (Alrubaian et al. 2018). Such features are derived from the user who has published the tweet, and assist with the third dimension of tweet credibility - how credible is the author of a tweet? Although we consider a user's network (e.g. number of followers and the number of accounts the user follows), other features are also considered; such as how long the user has been active on the Twitter platform (account age in days). As discussed in 7.2, research by Morris et al. (2012) has found that users that do not upload a custom profile picture, and instead use the default profile picture provided by Twitter, are perceived as less credible - a feature that is also considered in this methodology.

7.4.2 Financial Features

An overview of the FF group will be be discussed. As with GF, FF can also be divided into three groups: content, company-specific, and exchange-specific. The FF are the novel features that have yet to be considered within the literature as a means of assessing the credibility of stock tweets. It is anticipated that the inclusion of such features will contribute to improved performance when combined with the GF group. Many of the FF proposed depend on external sources that relate to a cashtag's corresponding company (e.g. the range of the company share price for that day). Features that are specific to the exchange are also proposed, such as: was the stock exchange open (i.e. actively trading) when the tweet was published. The FF groups will now be discussed further, starting with the content-based FF. The full list of FFs can be found in Table F.2, Appendix F.

[4] https://tweetdeck.twitter.com/

[5] https://www.hootsuite.com/en-gb/

[6] https://ifttt.com/

Content

Although numerous sentiment keyword lists exists for assessing the sentiment of a piece of text, certain terms are sometimes perceived differently in different contexts. For example, some keyword lists associate terms such as *death*, *mine*, and *drug* to be negative (Loughran and McDonald 2016), which means the use of such lists will lead tweets referring to companies that belong to the healthcare and mining sectors may be incorrectly be perceived as negative. Loughran and McDonald (2011) performed extensive research in establishing the sentiment of over 4,000 keywords in a financial context, and produced a keyword list (Table 7.1) that include groups such as positive, negative, and uncertainty keywords. Although other lexicons exist for the purpose of sentiment analysis on microblogging texts (Oliveira, Cortez, and Areal 2016), which may be effective, the lexicon produced by Loughran and McDonald (2011) was chosen due to its suitability for financial contexts. It should be noted, however, that there is an out-of-vocabulary issue to be aware of when using a set of keywords that are not strictly abiding of the Twitter way of communicating. Although abbreviating words is a very common practice on Twitter due to the character limit imposed upon tweets, keywords derived from formal financial documents (as is the case with Loughran and McDonald (ibid.)), are less likely to contain abbreviated communication speak. Word embedding would be helpful here, as this would allow similar words to have a similar encoding (e.g. allowing the abbreviated form of a word to considered the same as the formal full spelling). The count of words in each of these lists that is also found within the tweet text is transformed into its own respective feature when training the classifiers on the FF set.

Company-specific

Company-specific features are those that vary between stock-listed companies. Stock prices are provided in open, high, low, and close (OHLC) variants. These OHLC prices can pertain to a specific trading day, or a given time window (e.g. minutely, hourly). Two features are proposed that are engineering from the OHLC prices - the range of the high and low price for the day (Feature 50, Table F.2), and the range of the close and open price (Feature 51).

TABLE 7.1: Financial keyword groups as defined by (Loughran and McDonald 2011)

Keyword Group	Group Description	Total Keywords in Group	Keyword Examples
Positive	Positive in a financial context	354	booming, delighted, encouraged, excited, lucrative, meritorious, strong, winner
Negative	Negative in a financial context	2355	abnormal, aggravated, bankruptcy, bribe, challenging, defamation, disaster
Uncertainty	Indicates uncertainty	297	anomalous, could, fluctuation, probable, random
Litigious	Indicates litigious action (e.g. a lawsuit)	904	claimholder, testify, whistleblower, voided, ruling, perjury compel,
Constraining	Indicates constraints to a business	194	legal, employee, environmental, debt

Exchange-specific

Exchange-specific features are those that vary between stock exchanges. The count of credible financial hyperlinks in a tweet (Feature 54) requires the creation of a list of URLs that are deemed as being credible sources of information to that exchange. For example, London South East[7] is considered a reputable information source for stocks listed on the London Stock Exchange. Other features in this group include establishing if a tweet was published while the stock exchange was actively trading (stock exchanges have differently opening hours, are typically closed on the weekend, and some even take a lunch break in the middle of the day where trading ceases). With the feature groups introduced, the next section will discuss the feature selection techniques to be performed before, and during, the classifier training process.

[7] https://www.lse.co.uk/

7.5 Feature Selection

This section provides details on the different feature selection techniques that are proposed within the CAM (Section 7.3). As discussed in Section 7.2.1, much of the existing work on credibility does not focus on which features are most instructive when assigning credibility to tweets, this section will describe the different types of feature selection techniques that will result in the most informative feature set.

The aim of performing feature selection is to determine which inputs should be presented to a classification algorithm Omar et al. 2013. As previously discussed in Section 7.2.1, a large number of features may lead some machine learning algorithms to overfit, leading such algorithms to reach false conclusions and negatively affect their performance (Arauzo-Azofra, Aznarte, and Benítez 2011). Several other benefits of performing feature selection include improving interpretability and lowering the cost of data handling and acquisitions, thus improving the quality of such models. Some machine learning models have feature selection mechanisms embedded within them (referred to as embedded models). Decision trees, for example, have feature selection mechanisms embedded within them whereby the feature importance is calculated as the decrease in node impurity weighted b the probability of reaching that node (Ronaghan 2018). Naturally, Random Forest models also share this feature selection mechanism. Other machine learning models (e.g. Logistic Regression) often employ some kind of regularisation that punish model complexity by driving the learning process towards robust models by decreasing the less informative feature to zero and then dropping them (e.g. Logistic regression with L1-regularisation) (Richert 2013).

Feature selection techniques are typically classed as belonging to one of three groups: (1) filter methods, (2) wrapper methods, and (3) embedded methods. Each of these feature selection methods will now be discussed.

7.5.1 Filter Methods

Filter methods are often considered a pre-processing step before models are trained, in which the goal is to quickly screen the feature space to identify features that are, for example, constant, quasi-constant, or highly correlated. The benefit of this type of feature selection method is that it is undertaken before any models are trained, meaning they are computationally inexpensive and simply to perform (Tsai and Y.-C. Chen 2019). An extensive overview of different types of filter methods available for high-dimension classification data was recently undertaken by Bommert et al. (2020).

7.5.2 Wrapper Methods

Wrapper methods are another feature selection technique that aim to find the best subset of features according to a certain search strategy (Dorado et al. 2019). Wrapper methods involve the repeated training of classifiers on different feature sets to determine which features yield the best performance. Popular wrapped-based feature selection methods include sequential forward feature selection, sequential backward feature selection, and recursive feature elimination. As wrapper methods involve continuously re-training models on different feature sets, they do no scale particularly well to a large feature space.

7.5.3 Embedded Methods

Embedded feature selection methods incorporate the learning process of a classifier into the feature selection process (Hsu, Hsieh, and Lu 2011) and search for an optimal set of feature by optimising a function in advance. During the learning process, features that have little or no informative power are removed, meaning the features that have some predictive power remain in the final model. As is the case with wrapped methods, embedded methods are specific to the classifier being trained, with a key benefit being that embedded techniques communicate to the classifier, and are not as computationally expensive as wrapper methods (Rong, Gong, and X. Gao 2019).

The CAM (Fig 7.1) proposes using all three of these feature selection techniques. The filter methods are employed during the model preparation stage, with sequential-forward feature selection (wrapper method) using for each of the classifiers. Models that feature embedded feature selection techniques within them (e.g. decision tree, random forest, logistic regression) inherently perform embedded feature selection due to the nature of these algorithms. The next section will discuss the experimental design to validate the CAM methodology presented in Section 7.3.

7.6 Experimental Design

To validate the CAM (Section 7.3), an experiment was designed involving companies listed on the London Stock Exchange (LSE). This section will discuss the experiment, details on how the dataset was created, how the annotation of the dataset was performed, and list the most informative features as a result of performing the feature selection techniques discussed in Section 7.5

7.6.1 Company Selection

The credibility classifiers considered tweets corresponding to 100 companies listed on the London Stock Exchange (Appendix E). Companies were chosen from each of the different industries defined by the LSE (e.g. oil & gas, telecommunications, financial services) and companies that had not been listed on the LSE for at least two years were excluded from being shortlisted in order to maximise data collection.

7.6.2 Data Collection

Tweets containing at least one occurrence of a cashtag corresponding to the ticker of at least one of the companies listed in Appendix E were collected across a 1-year period (15/11/19 - 15/11/20. In total, 208,209 tweets were collected over the one-year period. Numerous FFs require data from various APIs (e.g. share prices). Daily share prices spanning the same time period were collected by AlphaVantage. Broker ratings and dates in which Regulatory News Service (RNS) announcements were made have been web scraped from London South East [8] (as several of the FF include considering the number of broker ratings and RNSs issued on the day of the tweet).

7.6.3 Tweet Annotation

After tweets containing at least one occurrence of a cashtag of a company in Appendix B, a subsample of 5,000 tweets were shortlisted to be annotated. We began

[8] lse.co.uk

by attempting to select 25 tweets for each of the experiment companies (as listed in Appendix E), which resulted in 3,874 tweets - we then randomly select tweets to reach a total of 5,000.

As discussed in Section 7.2, credibility is subjective, and annotating datasets for credibility is likely to vary significantly between different annotators depending on their perceptions and experiences. If a subjective-type dataset is annotated by a single individual, then it will result in classifiers that have learned the idiosyncrasies of that particular annotator (Reidsma and Akker 2008). In order to alleviate this issue, we began by having a single annotator (referred herein to as the main annotator - MA) provide labels for each tweet based on a five-label Likert scale (Joshi et al. 2015) system (Table 7.5). A subset of these tweets (10) was then selected and shown to three other annotators (annotators 1-3 - A1, A2, A3) who have had previous experience with Twitter datasets, to establish the inter-item correlation between the annotators' annotations. In order to assess the inter-item correlation between the different annotators, the Cronbach Alpha (CA) score (Equation 7.1) was obtained for the different annotation scenarios.

$$\alpha = \frac{N\overline{c}}{\overline{v} + (N-1)\overline{c}} \tag{7.1}$$

where N is the number of items, $\bar{c}$ is the average inter-item covariance among the items and $\bar{v}$ is the average variance. A CA score of more than 0.7 implies a high level of agreement between the annotators (Landis and Koch 1977). The CA for the binary labelled tweets (Table 7.2) is 0.591 - indicating the annotators were unable to reach a consensus on deciding if a tweet was credible or not. The CA for the five-label system was then computed (Table 7.3), in which the CA was 0.699. The CA for the five-label system shows that the annotators were able to find a more consistent agreement, but did not meet the threshold that is considered a high level of agreement. An additional experiment involving a three-label system (not credible, ambiguous, and credible), with a larger sample size of 30 tweets was then undertaken to assess the annotators' agreement on such a scale. In each of these three experiments, it is clear that is the CA is computer with the MA annotations removed, it will result in the greatest decrease in the CA score. This indicates that the majority of the annotators' annotations are mostly aligned with the MA. Although none of these experiments led to a CA of more than 0.7 (the threshold constituting a high agreement), we seek to find a consensus between the majority of annotators - as long as the MA is not in a minority. The higher CA score (from the majority - 3) comes from using the binary-labelled system where the annotators annotated tweets as being credible or not credible, in which the CA becomes 0.895 if the first annotator's (A1) annotations are removed. In other words, the MA, annotator 2 (A2), and annotator 3 (A3) were able to reach a consensus on annotating credibility when using a binary annotation approach. However, a binary approach does not provide a lot of granularity when compared to a multi-class approach. Due to the five-class system having a significant class imbalance when taking into consideration the individual classes (814 strong not credible vs 1320 not credible tweets), a three-class system that combines the two not-credible classes and the two credible classes is used to ensure that ambiguous tweets can be taken into consideration (Table 7.4).

TABLE 7.2: Inter-Item Correlation Matrix & CA Scores for binary-labelled tweets. CA = 0.591 (Sample size = 10)

	MA	A1	A2	A3	CA if item deleted
MA	1.000	-0.200	0.816	0.816	0.148
A1	-0.200	1.000	0.000	-0.408	0.895
A2	0.816	0.000	1.000	0.583	0.179
A3	0.816	-0.408	0.583	1.000	0.433

TABLE 7.3: Inter-Item Correlation Matrix & CA Scores for five-class labelled tweets. CA = 0.699 (Sample size = 10)

	MA	A1	A2	A3	CA if item deleted
MA	1.000	-0.061	0.722	0.827	0.443
A1	-0.061	1.000	0.210	-0.063	0.866
A2	0.722	0.210	1.000	0.578	0.538
A3	0.827	-0.063	0.578	1.000	0.518

7.6.4 Assessing Feature Importance

As discussed in Section 7.5, by assessing the informative power of the features discussed in Appendix F, features that do not offer any benefit to classifiers can be removed so that more robust classifiers can be trained to assess the credibility of

TABLE 7.4: Inter-Item Correlation Matrix & CA Scores for three-class labelled tweets. CA = 0.686 (Sample size = 30)

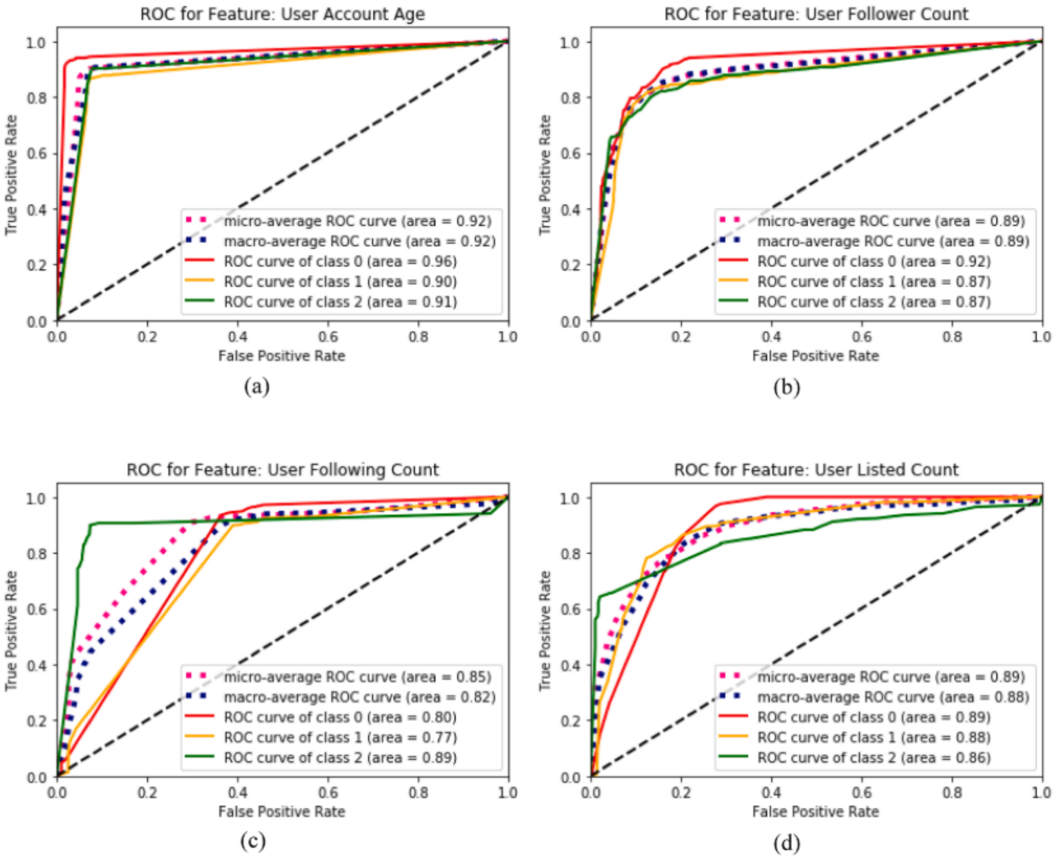
	MA	A1	A2	A3	CA if item deleted
MA	1.000	0.715	0.752	0.173	0.449
A1	0.715	1.000	0.600	0.052	0.547
A2	0.752	0.600	1.000	0.055	0.537
A3	0.173	0.052	0.055	1.000	0.866

TABLE 7.5: Annotated tweet breakdown

Label	Description	Tweet Count	Merged Count
0	Strong Not Credible	814	
1	Not Credible	1320	2134
2	Ambiguous / not enough information	693	693
3	Fairly credible	1020	
4	Very credible	1153	2173

financial stock tweets. To assess the informative power of each feature, a Decision Tree (DT) classifier has been trained on each feature, to assess the informative power of the feature on its own.

The metric used to calculate the importance of each feature is the probability returned from the DT. We then calculate the total area under the curve (AUC) for the feature. Naturally, the AUC can only be computed for a binary classification problem. In order to calculate the AUC for a multi-class problem, the DT classifier, which is capable of producing an output y = 0, 1, 2 (matching the three-level annotation system), is converted into three binary classifiers by adopting a One-Vs-Rest approach (Ambusaidi et al. 2016). Each of the AUC for the three binary classifier (for each feature), is then calculated to establish the feature importance for each class. The AUC score can be computed in several ways for a multiclass classifier. The macro average computes the metric for each class independently before taking the average, whereas the micro average is the traditional mean for all samples (Aghdam, Ghasem-Aghaee, and Basiri 2009). Macro-averaging treats all classes equally, whereas micro-averaging favours majority classes. We elect to judge the informative power of the feature based on its AUC macro average, as ambiguous tweets are relatively more uncommon than credible and not credible tweets in our dataset. The four most informative features (based on the macro AUC score) are depicted in Figure 7.3, each of these posses an AUC score of more than 0.8 - indicating that these



(a)

(b)

(c)

(d)

FIGURE 7.3: Top four informative features based on macro AUC

features are highly informative. These four features are all contained within the general group, and is consistent with previous work that has found user-based features to be particularly powerful for assessing credibility (F. Yang et al. 2012). The filter methods outlined in the CAM (Figure 7.3) has been applied to the feature set of the 5,000 annotated tweets to highlight features that will not offer any benefit to classifiers. Based on the five filter method feature selection techniques outlined in Table 7.6, 18 features were identified as not offering any meaningful informative power to the classifiers (based on the probability returned from the DT). Now that the informative and non-informative features have been identified, classifiers can be trained on an optimal feature set. The 18 features identified in Table 7.6 have been dropped due to the reasons outlined in Table 7.6.

TABLE 7.6: Features removed via feature selection techniques

Feature Selection Technique	Description	Features Identified
Constant features	Features that are constant across all observations	Tweet contains positive emoticons Tweet contains negative emoticons
Quasi-constant features	Features that are constant across almost all observations	Tweet contains multiple question marks Tweet contains exlamation mark Count of second-person pronouns User is verfied Tweet is a quote tweet Tweet contains media Interjection word count Count of constrinaing keywords
Duplicated features	Features that convery the same information	*None*
Highly-correlated features	Features with a Pearson's correlation coefficient of >0.8	User has non-fictional location Tweet is a retweet Tweet length (words) Username word count
Univariate ROC-AUC score	Features that have a ROC-AUC score close to random chance	Financial cashtags in tweet Technology cashtags in tweet Telecommunication cashtags in tweet

7.7 Results & Discussion

This section will report the results (Table 7.7) of training different supervised classifiers for assessing the credibility of stock tweets after the non-informative features were removed, and illustrate how the performance of some classifiers decreases if feature selection techniques are not adopted. Classifiers that have previously enjoyed success in assessing the credibility of microblog posts have been trained (Alrubaian et al. 2018). These include Naive Bayes, k-Nearest Neighbours, Decision Trees, Logistic Regression, and Random Forest). The results presented are based on 10-fold cross validation using an 80/20 train-test split and have been implemented using the sci-kit learn library in Python. Each of the classifiers underwent a hyperparameter grid search to seek out the most optimal hyperparmeters. Three groups of classifiers have been trained: (1) trained on the general features, (2) trained on the financial features, and (3) trained on both general and financial features.

The wrapper feature selection technique, Sequential Forward Feature Selection

(SFFS), found that the kNN and NB classifiers suffered significant performance decreases as the feature set for those classifiers grew (depicted in Figure 7.4). This is due to the well-documented phenomenon of the curse of dimensionality (Parmezan, H. D. Lee, and Wu 2017). Classifiers that have natural (embedded) feature selection mechanisms built-in to them were robust to the size of the feature space increasing (DT, RF, and LR). Based on the ROC-AUC, the RF classifier was the best-performing when trained on the combined feature set of general features (GF) and financial features (FF). Classifiers that only considered the FF paled in comparison to classifiers trained on the general features - meaning the FF alone are not suitable for assessing credibility of stock tweets. As indicated by the results, classifiers trained on the GFs perform very well, with slight increases in the AUC when FFs were added to the GFs.

The importance of feature selection is evident from the SFFS performed, particularly for the kNN classifier, which reaches its performance peak at 9 features, and almost outperforms the RF when both are compared at the same number of features. In respect to the five classifiers trained on the combined features, the most popular FFs utilised by the classifiers were the count of cashtags in the tweet (F58), and the count of technology and healthcare cashtags within the tweet (2xF59+).

As evident from the initial experiment results in Table 7.7, RF appears to be the best performing classifier when trained on both GF and FF. We now test to see if the differences between the test set predictions of the RF trained on GF versus the RF trained on the combined features is statistically significant by conducting the Stuart-Maxwell test. The Stuart-Maxwell test is an extension to the McNemar test, used to assess marginal homogeneity in independent matched-pair data, where responses are allowed more than two response categories (Z. Yang, Sun, and Hardin 2011). The p-value of the Stuart-Maxwell test on the predictions of both the RF trained on GF and the RF trained on the combined features is 0.0031, indicating the difference between the two classifiers is statistically significant.

TABLE 7.7: Credibility classifier results

	General Features						Financial Features						General + Financial Features					
Classifier	Features (/34)	Acc	Pre	Rec	F1	AUC	Features (/21)	Acc	Pre	Rec	F1	AU C	Features (/55)	Acc	Pre	Rec	F1	AUC
NB	4	85.5	84.8	85.5	85.0	89.1	12	61.0	63.9	60.3	59.7	70.4	6 (2FF)	85.6	84.9	85.6	85.1	91.4
LR	21	88.0	84.6	86.0	85.3	90.5	9	55.9	40.8	50.7	43.0	64.0	27 (9FF)	87.6	87.1	86.8	86.9	92.0
DT	18	90.1	90.6	90.4	90.5	92.6	10	54.2	55.1	49.6	43.0	63.1	11 (3FF)	89.7	90.1	90.0	90.0	93.1
RF	20	92.7	93.1	92.6	92.9	93.8	11	61.9	63.1	60.9	60.4	70.9	37 (12FF)	**93.5**	**94.3**	**93.2**	**93.7**	**94.3**
kNN	7	91.4	92.3	91.1	91.6	93.2	7	61.5	64.0	61.3	60.8	71.1	9 (2FF)	92.7	93.6	92.5	92.9	93.6

Note: Scores presented are the macro average percentage (%).

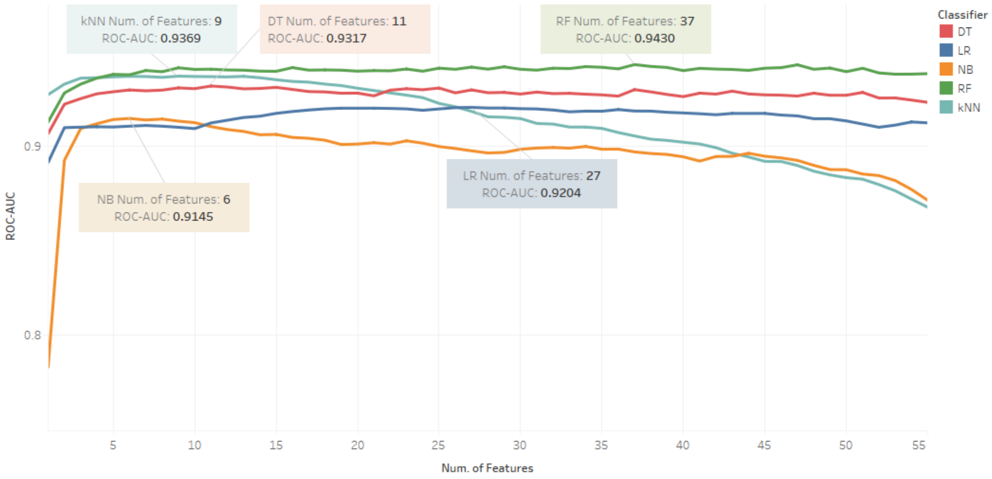


FIGURE 7.4: Sequential forward feature selection results (combined feature set)

7.8 Chapter Summary

This chapter has presented a methodology for assessing the credibility of financial stock tweets. Two groups of features were proposed: (1) general features that can be derived from any tweet, and (2) financial features that can be created from financial stock tweets (tweets containing a cashtag). Feature selection techniques were utilised before classifiers were trained to identify features that would offer little or not informative power to classifiers. Three sets of classifiers were trained, taking into consideration general features, financial features, and a combination of the two. Performance gains were obtained by combining the two groups of features in the training of the classifiers, with NB and kNN classifiers suffering performance decreases when the groups of features were combined.

Although the RF classifiers (trained on GF or both GF and FF) were certainly the best performing classifiers in respect to the AUC, the kNN classifier trained on the combined feature set was also a formidable classifier due to its comparative performance with the RF classifiers without having to take into account as many features (9 features for kNN compared to 37 for RF). The Random classifier was persisted and deployed to the SDE data layer to assess the credibility of stock tweets during the fusion process.

The main contributions of this chapter can be summarised as follows:

- A novel methodology for assessing the credibility of financial stock tweets, trained on a novel set of features
- The importance of feature selection for assessing financial stock tweets is highlighted, particularly when considering classifiers that suffer decreased performance as the feature space increases

Chapter 8

Detection Layer: Detection of Financial Market Irregularities

8.1 Overview

This chapter presents the detection layer of the Smart Data Ecosystem (SDE) introduced in Section 4.2. The detection layer (Figure 8.1) focuses on detecting irregularities by honing in on specific company time periods referred to as *events* (defined and discussed in Section 8.3). Much of the previous work that attempts to detect irregular activity focuses on events, such as a company making an announcement, a merger or acquisition (Fernández Vilas et al. 2017), the appointment of a new CEO (Gondhalekar and Dalmia 2007; Byrka-Kita, Czerwiński, and Preś-Perepeczo 2017), or the effects that broker ratings have on share prices (Sabherwal, Sarkar, and Y. Zhang 2011).

This chapter begins by introducing the methodology to generate events (Section 8.3) for the detection layer to focus on. A high-level overview of the detection layer methodology is then provided in Section 8.4. The first type of clustering supported by the detection layer - the clustering of events - is then presented in Section 8.5. The clustering of tweets within an event is then discussed in Section 8.6. The clustering of FDB posts within an event is then presented in Section 8.7. The detection mechanisms presented in this chapter are evaluated through qualitative interviews conducted with five financial market experts, the results of the evaluation are presented in Chapter 9. This chapter aims to specifically address the second and third research questions outlined in Section 1.5 - *Can a smart data ecosystem be used to monitor a variety of communication channels for irregular behaviour?*, and *Can a smart data ecosystem, utilising clustering algorithms, identify irregular days and events with respect to posting activity?*

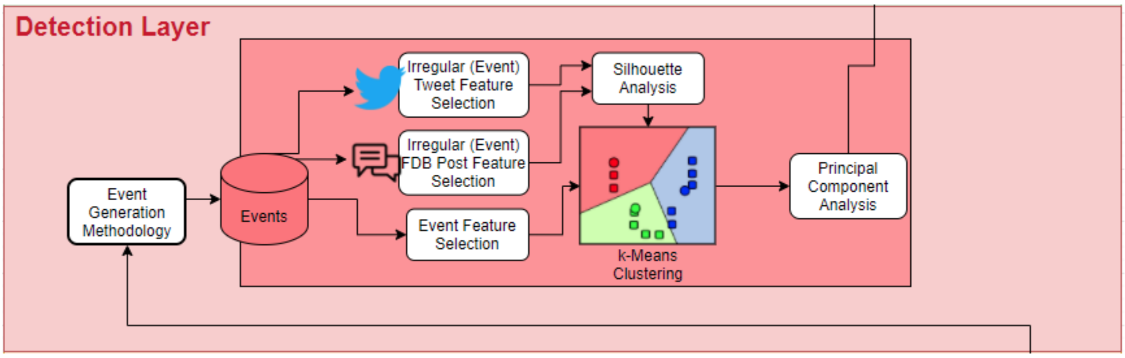

Detection Layer
Event Generation Methodology
Events
Irregular (Event) Tweet Feature Selection
Irregular (Event) FDB Post Feature Selection
Event Feature Selection
Silhouette Analysis
k-Means Clustering
Principal Component Analysis

FIGURE 8.1: SDE Detection Layer

8.2 Irregularity Detection via clustering

Unsupervised clustering algorithms have played an important role in the detection of irregularities. Xu and Tian (2015) state that traditional clustering algorithms can be divided into nine categories, which primarily consist of 26 algorithms. The most adopted clustering algorithm being k-means, which clusters a data point to a k group (centroid) based on a pre-defined distance metric through iterative computation. The crucial step in this clustering algorithm is establishing the number of clusters (k), although techniques such as silhouette analysis can be of assistance in this regard (Géron 2019).

Over the years, variants of k-means algorithms have been proposed, including: continuous k-means (Faber 1994), trimmed k-means (Cuesta-Albertos, Gordaliza, and Matrán 1997), compare means (Phillips 2002), k-probabilities (Wishart 2003), X-means(Pelleg and Moore 2015), k-modes (Chaturvedi et al. 2001), k-harmonic (L. D. Zhang et al. 2013) and k-prototype (Z. Huang 1998). The basic k-means algorithm, however, has remained steadfast and continues to be the dominant unsupervised clustering algorithm (M. Ahmed, Mahmood, and Islam 2016).

8.3 Event Generation Methodology

The detection layer event generation methodology (Figure 8.2) is responsible for generating *event* documents that correspond to company events. In the context of this research, an event is defined as a significant moment in a company's operations - for example, a company's CEO stepping down, a broker agency revising its rating for a company (buy/sell ratings), or a company publishing an RNS announcement to address speculation or rumour regarding its operations. An event document stores all data pertinent to an event spanning a two-week window as specified in Figure 8.2, this includes all data retrieved from the time-slice windows (i.e the event document contains all of the time-slice windows for the two-week period), and summary data such as how many credible tweets were made in the week leading up to the event, and . Events are stored in a NoSQL events database and contain all of the data within an *event window*, such as the tweets, FDB posts, and share prices for each day in the event window. Based on empirical analysis, an event window spans a two-week period which comprises a pre-event window and a post-event window. The pre-event window is seven days before the event occurred (e.g. a buy/sell broker rating being issued), with the post-even window being seven days immediately after the event occurred.

In order to provide a proof-of-concept, the SDE generates an *event document* whenever a buy or sell broker rating is issued for one of the SDE companies in Appendix B. Buy and sell broker ratings have been selected due to many companies within the SDE possessing broker ratings, versus rarer triggers such as a CEO of a company changing. Buy and sell broker ratings also recommend taking a course of action to investors (the buying or selling of shares), which could provide insights into investor behaviour during the broker rating period. The next section will provide an overview of the irregularity detection methodology.

8.4 Irregularity Detection Methodology

Once events have been generated using the methodology outlined in Section 8.3, the clustering process can begin. The detection layer supports three types of clustering:

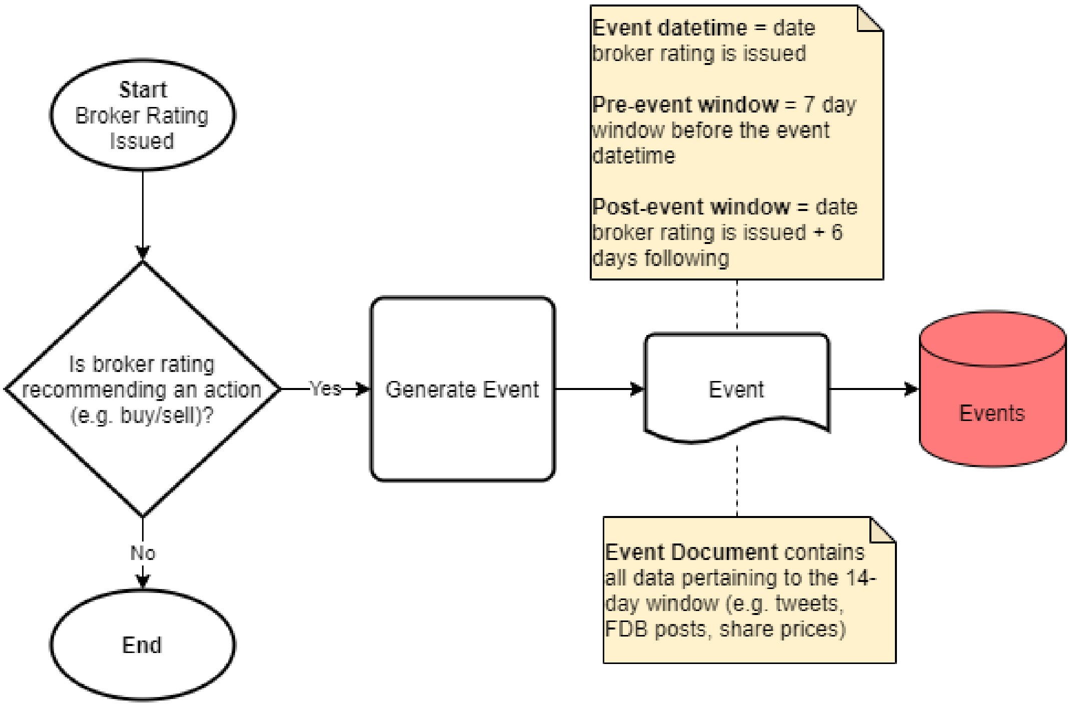


FIGURE 8.2: Event generation methodology

1. **Clustering of events** - Events (two-week time periods) for a single company can be clustered (Section 8.5). Events are clustered based on the posting (tweets and FDB posts) activity, the breakdown of tweets in respect to their credibility, and the number of unique twitter and discussion board users participating in the discussion.

2. **Clustering of tweets within an event** - All tweets within an event can be clustered to identify potentially irregular tweets (Section 8.6).

3. **Clustering of financial discussion board posts within an event** - All FDB posts within an event can be clustered to identify potentially irregular posts (Section 8.7).

The full list of features used in each of these three types of clustering can be found in Appendix G.

The features used for clustering events originated by focusing on two aspects of an event: the pre-event window, and the post-event window. The pre-event window features focus on the one week period before the event occurred, with the post-event window features focusing on the one week period after the event occurred. This division of features was chosen to aid the interpretation for users after the clustering has been accomplished - i.e. are data points more likely to belong to an outlying cluster if their pre-event discussion features (e.g. pre-event credible tweets) are more abnormal than their post-event features? These features were extracted by looking at the date and timestamps of the various data sources - i.e. if a broker rating (event) occurred at 8am on a Monday, the pre-event window would be derived by looking at all data points from the previous Monday at 8am, and the post-event window would be derived by looking at all data points up to the next Monday at 8am.

The features used for clustering FDB posts focused on the different attributes collected for the FDB posts, e.g. is the user posting a premium member? The stock price of the stock at the time of the post (this may help in identifying outliers, as many of the posts made could have been posted when the stock price was high, meaning minority posts made when the stock price was low could be of interest).

The features used for clustering tweets relate primarily to the metadata found within the tweet JSON object, e.g. the number of cashtags, hashtags, media (images/videos) in the tweet. It may be typical for tweets for a specific company to only contain a couple of cashtags, meaning tweets containing many more than this would be considered irregular and belong to its own outlying cluster.

It is important to mention at this point that if non-LSE tweets were not removed (using the cashtag collision methodology discussed in Section 6.6), then the clustering of tweets would be a fruitless task, as cryptocurrency and non-LSE stock tweets would undoubtedly impact the clustering process and make any analyses of tweets difficult to undertake. This is primarily due to the different characteristics of such tweet groups - cryptocurrency tweets typically have a higher number of cashtags within them, meaning a feature engineered from the cashtag count will naturally lead to cryptocurrency tweets being clustered together.

8.4.1 Choosing the optimal number of clusters

One of the principal challenges with using unsupervised clustering algorithms such as k-means clustering is choosing the optimal number of clusters - the k value. Several techniques exist for selecting an optimal k value for the k-means clustering algorithm, including the elbow method (Bholowalia and A. Kumar 2014), information criterion approach (Bozdogan 1987), and silhouette analysis (Rousseeuw 1987).

Although the elbow method is a well-established method of choosing an optimal *k* value for clustering, it is a manual method that requires inspecting the within-cluster sum-of-square (WCSS) as plotted on a graph. Techniques such as silhouette analysis, on the other hand, can be used to study the separation distance between the different clusters and results in a silhouette score being returned for every *k* value in a range. The silhouette score ranges from -1 to +1, with higher values indicating that an object is well matched to its own cluster and poorly matched to neighbouring clusters (Zhou and J. T. Gao 2014).

When selecting the number of clusters for clustering the events, the default *k* value is set to 2 in order to provide two groupings - regular events and irregular events. This *k* value can be overridden within the GUI if two clusters does not provide a clear group of outliers. If a *k* value is not specified, multiple k-means models are trained, and the model with the highest silhouette score is used when providing a visual representation of the clusters.

When selecting the number of clusters for clustering tweets or FDB posts within an event, the detection layer will train multiple k-means clustering models (with (*k* values from 2 to 10), and return the best-performing model in respect to the silhouette score. As with the event clustering, the *k* value can also be manually specified with a user-specific *k* value.

8.4.2 Visualising the results via Principal Component Analysis (PCA)

Once the clustering process has concluded, the popular dimensionality-reduction algorithm, Principal Component Analysis (PCA), proposed by Wold, Esbensen, and Geladi (1987), is utilised to assist in interpreting the clustering results. PCA involves the creation of new variables (referred to as the *principal components)* which are computed by extracting the most important information from the given feature set (Tsai and Y.-C. Chen 2019). The new variables that are created are linear combinations of the original variables, with the top principal components explaining the largest variance within the dataset. In other words, if two principal components are created, the first principal component will explain more of the variance than the second principal component. PCA is often used prior to clustering in order to reduce the number of features to a number that is easier to visualise (2-3), it is also thought to reduce noise in the data, potentially leading to improved results when performing the clustering process (Kaloyanova 2020). Although other dimensionality reduction algorithms exist, due to the success enjoyed in the literature (Avalon et al. 2017) in combining these approaches, PCA was chosen as the dimensionality reduction algorithm for the detection layer. By extracting two principal components from the feature set, these can be used to plot the data points (and their associated cluster) on a two-dimensional space for easier interpretation.

8.5 Irregular Event Detection

The first type of clustering the detection layer can perform is considered a high-level clustering in which the two-week event windows for a company are clustered based features such as the volume of tweets, FDB posts, and the breakdown of credibility for tweets in the event window. The features used to cluster events are presented in Appendix G (Table G.1). The features are split into two groups: (1) pre-event features and (2) post-event features.

An example of the event clustering is shown in Fig 8.3. In this example, ten events (buy or sell broker ratings) are visually represented on the graph. Nine of these data points have been clustered into cluster 1, with one data point assigned to cluster 2.

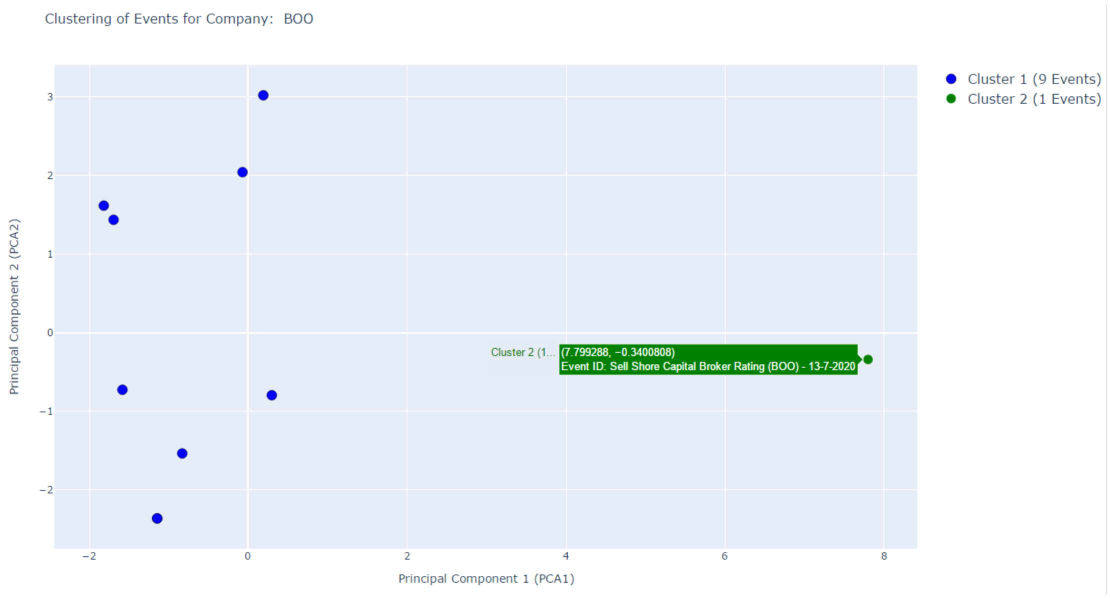


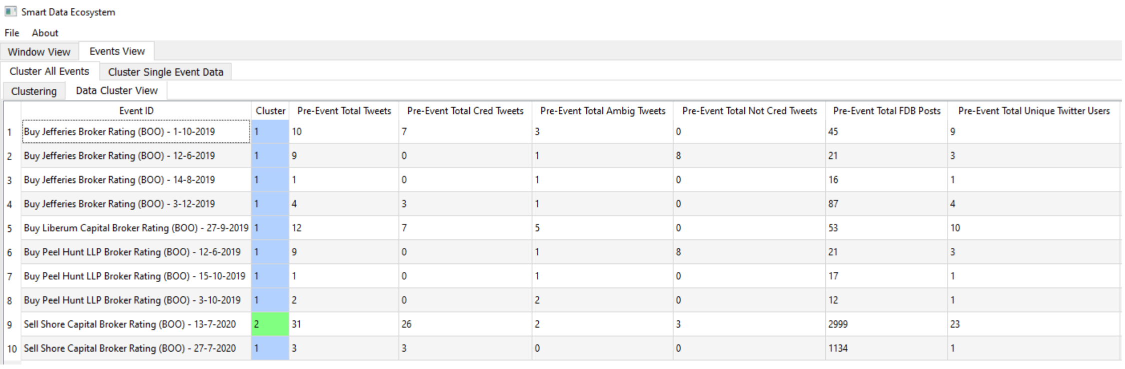
FIGURE 8.3: Clustering of events for Boohoo PLC (LON:BOO)

Once events for a company have been clustered, the individual data points (individual events for that company) can be viewed via the **Data Cluster View** tab (depicted in Figure 8.4). This view shows the individual data points and the features used to cluster the data points, which aids the interpretation of the clusters and their respective data points. The single data point assigned to cluster 2, for example, can be inspected further to establish why it may have been assigned to its own cluster (i.e. analysing and comparing the pre and post-event features with data points in other clusters).

Specifically in the example depicted in Fig 8.3 and Fig 8.4, the single data point assigned to the second cluster (green) can be observed to have significantly higher counts of tweets within the pre-event window, and a dramatic increase in FDB posts in the pre-event window (2999 FDB posts, whereas other events for this company typically have less than 100 FDB posts in the pre-event period). A detailed analysis of this clustering functionality is provided in Section 9.5.4, in which five financial market experts review the SDEs event clustering capabilities.

8.6 Irregular Tweet Detection

Irregular tweets within an event can also be clustered by the detection layer. Firstly, an event must be selected for a company. Then, all of the tweets made within that event window can be clustered. The default behaviour of this clustering type is to train multiple k-means models using different k values for each model (where the k value ranges from 2 to 10 inclusive). The silhouette score for each of these models is compared, and the best-performing model (with respect to its silhouette



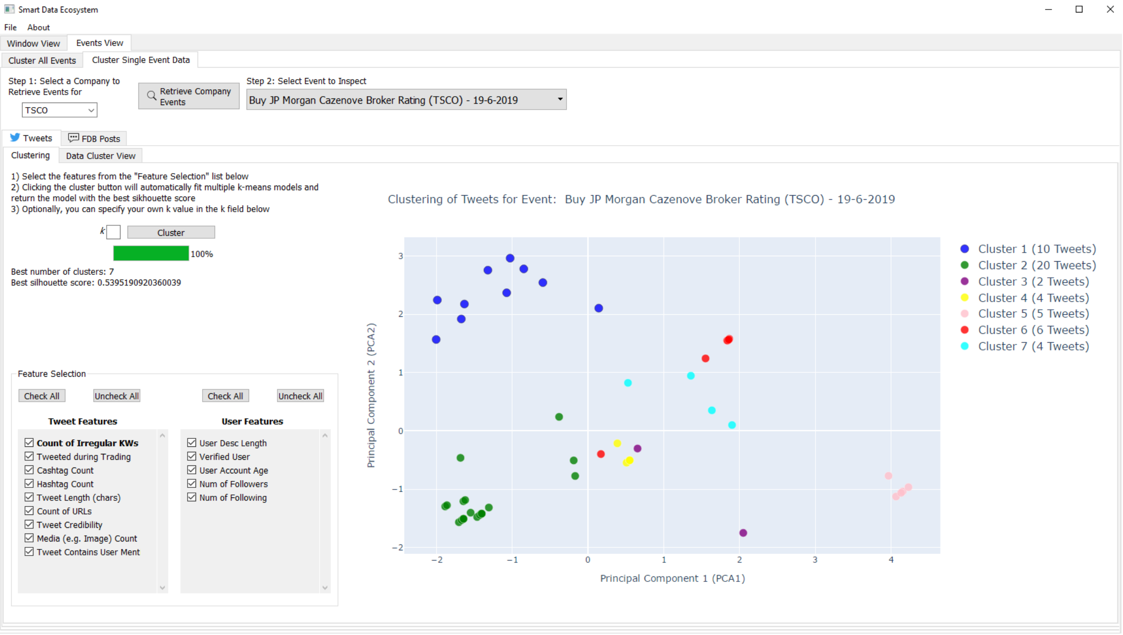
	Event ID	Cluster	Pre-Event Total Tweets	Pre-Event Total Cred Tweets	Pre-Event Total Ambig Tweets	Pre-Event Total Not Cred Tweets	Pre-Event Total FDB Posts	Pre-Event Total Unique Twitter Users
1	Buy Jefferies Broker Rating (BOO) - 1-10-2019	1	10	7	3	0	45	9
2	Buy Jefferies Broker Rating (BOO) - 12-6-2019	1	9	0	1	8	21	3
3	Buy Jefferies Broker Rating (BOO) - 14-8-2019	1	1	0	1	0	16	1
4	Buy Jefferies Broker Rating (BOO) - 3-12-2019	1	4	3	1	0	87	4
5	Buy Liberum Capital Broker Rating (BOO) - 27-9-2019	1	12	7	5	0	53	10
6	Buy Peel Hunt LLP Broker Rating (BOO) - 12-6-2019	1	9	0	1	8	21	3
7	Buy Peel Hunt LLP Broker Rating (BOO) - 15-10-2019	1	1	0	1	0	17	1
8	Buy Peel Hunt LLP Broker Rating (BOO) - 3-10-2019	1	2	0	2	0	12	1
9	Sell Shore Capital Broker Rating (BOO) - 13-7-2020	2	31	26	2	3	2999	23
10	Sell Shore Capital Broker Rating (BOO) - 27-7-2020	1	3	3	0	0	1134	1

FIGURE 8.4: Data points view of Boohoo PLC (LON:BOO) events

score) is plotted on a two-dimension space using the PCA technique discussed in Section 8.4.1. This default behaviour can be overridden by specifying a k value in the relevant field on the GUI.

An example of the tweet clustering process is depicted in Fig 8.5. Each of the data points represents a tweet (classified as belonging to the LSE by the classifier discussed in Section 6.6) within the event window. As with the event clustering, the individual data points of this clustering type can be compared using the **Data Cluster View** tab (Fig 8.6).

In the example depicted in Fig 8.5, the most populous cluster is cluster 2 containing 20 tweets - containing twice as many tweets are the next populous cluster (cluster 1 with 10 tweets). Upon closer inspection of the data points in cluster 1, all but one of the tweets were classified as being not credible by the SDEs credibility classifier (Section 7.8). The remaining data points belonging to clusters 2-7 were either credible or ambiguous tweets - indicating that the credibility of tweets can often be used as a means of determining a data point's cluster when considered as a feature.



Step 1: Select a Company to Retrieve Events for: TSCO

Step 2: Select Event to Inspect: Buy JP Morgan Cazenove Broker Rating (TSCO) - 19-6-2019

1) Select the features from the "Feature Selection" list below
2) Clicking the cluster button will automatically fit multiple k-means models and return the model with the best sikhouette score
3) Optionally, you can specify your own k value in the k field below

k □ Cluster

100%

Best number of clusters: 7
Best silhouette score: 0.5395190920360039

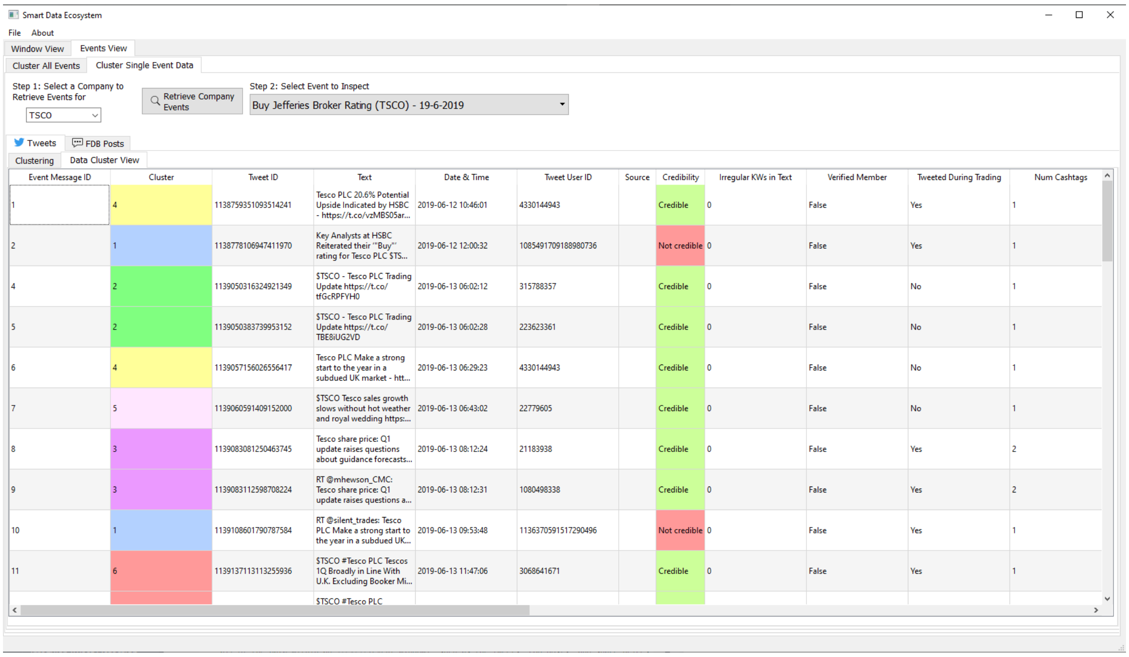
Tweet Features

☑ **Count of Irregular KWs**
☑ Tweeted during Trading
☑ Cashtag Count
☑ Hashtag Count
☑ Tweet Length (chars)
☑ Count of URLs
☑ Tweet Credibility
☑ Media (e.g. Image) Count
☑ Tweet Contains User Ment

User Features

☑ User Desc Length
☑ Verified User
☑ User Account Age
☑ Num of Followers
☑ Num of Following

FIGURE 8.5: Clustering of tweets for Tesco (Event: Buy rating by JP Morgan 19/06/2019)



Step 1: Select a Company to Retrieve Events for: TSCO

Step 2: Select Event to Inspect: Buy Jefferies Broker Rating (TSCO) - 19-6-2019

Event Message ID	Cluster	Tweet ID	Text	Date & Time	Tweet User ID	Source	Credibility	Irregular KWs in Text	Verified Member	Tweeted During Trading	Num Cashtags
1	4	1138759351093514241	Tesco PLC 20.6% Potential Upside Indicated by HSBC - https://t.co/vzMBS05ar...	2019-06-12 10:46:01	4330144943		Credible	0	False	Yes	1
2	1	1138778106947411970	Key Analysts at HSBC Reiterated their '"Buy"' rating for Tesco PLC $TS...	2019-06-12 12:00:32	1085491709188980736		Not credible	0	False	Yes	1
4	2	1139050316324921349	$TSCO - Tesco PLC Trading Update https://t.co/tfGcRPFYH0	2019-06-13 06:02:12	315788357		Credible	0	False	No	1
5	2	1139050383739953152	$TSCO - Tesco PLC Trading Update https://t.co/TBE8iUG2VD	2019-06-13 06:02:28	223623361		Credible	0	False	No	1
6	4	1139057156026556417	Tesco PLC Make a strong start to the year in a subdued UK market - htt...	2019-06-13 06:29:23	4330144943		Credible	0	False	No	1
7	5	1139060591409152000	$TSCO Tesco sales growth slows without hot weather and royal wedding https:...	2019-06-13 06:43:02	22779605		Credible	0	False	No	1
8	3	1139083081250463745	Tesco share price: Q1 update raises questions about guidance forecasts...	2019-06-13 08:12:24	21183938		Credible	0	False	Yes	2
9	3	1139083112598708224	RT @mhewson_CMC: Tesco share price: Q1 update raises questions a...	2019-06-13 08:12:31	1080498338		Credible	0	False	Yes	2
10	1	1139108601790787584	RT @silent_trades: Tesco PLC Make a strong start to the year in a subdued UK...	2019-06-13 09:53:48	1136370591517290496		Not credible	0	False	Yes	1
11	6	1139137113113255936	$TSCO #Tesco PLC Tescos 1Q Broadly in Line With U.K. Excluding Booker Mi...	2019-06-13 11:47:06	3068641671		Credible	0	False	Yes	1
			$TSCO #Tesco PLC								

FIGURE 8.6: Tweets data points view (LON:TSCO) 19/06/2019

8.7 Irregular FDB Post Detection

The detection layer also facilitates the clustering of FDB posts within an event to identify potentially irregular posting activity. An example of the FDB clustering is depicted in Figure 8.7. This example shows 5,786 posts being clustered for an event pertaining to Boohoo PLC (LON:BOO). The most interesting cluster could be perceived to be the cluster that contains the least data points (FDB posts) assigned to it - in this case, cluster 6 with seven data points. When inspecting the data points of cluster 6, they all share a commonality that none of the other data points in the other clusters do - they all contain keywords that could indicate the user of the FDB posts has some kind of insider knowledge.

When visualising the clusters after PCA has been performed, there is some visual overlap in some data points seeming to overlap with data points from other clusters, this is a natural side-effect of the PCA algorithm not being able to capture 100% of the underlying variance of all of the features it is attempting to summarise. In other words, the PCA components (x-axis being PCA1, and y-axis being PCA2) are based on the two principal components being plotted on their respective axes, whereas the cluster a data point is assigned to is based on all of the features used in the clustering process, meaning several data points appear to 'belong' to neighbouring clusters when visualised using the two principal components.

One of the features used to cluster FDB posts and tweets is the *number of keywords of interest*. This is a feature that is derived from existing research (Owda, Crockett, and P. S. Lee 2017) and keywords obtained by experts. The keywords of interest detected in the posts in cluster 6 include the phrase "told me" 8.8. The first post in this cluster (Event Message ID: 1131) contains the text: "I know someonme who work sin the warehouse in Burnley and they told me yesterday there was 180k items being processed at the time they [...]". Although this post contains typographical errors, the poster claims to have some knowledge which may not be considered in the public domain - which could constitute market abuse according to the Financial Conduct Authority (Financial Conduct Authority 2021).

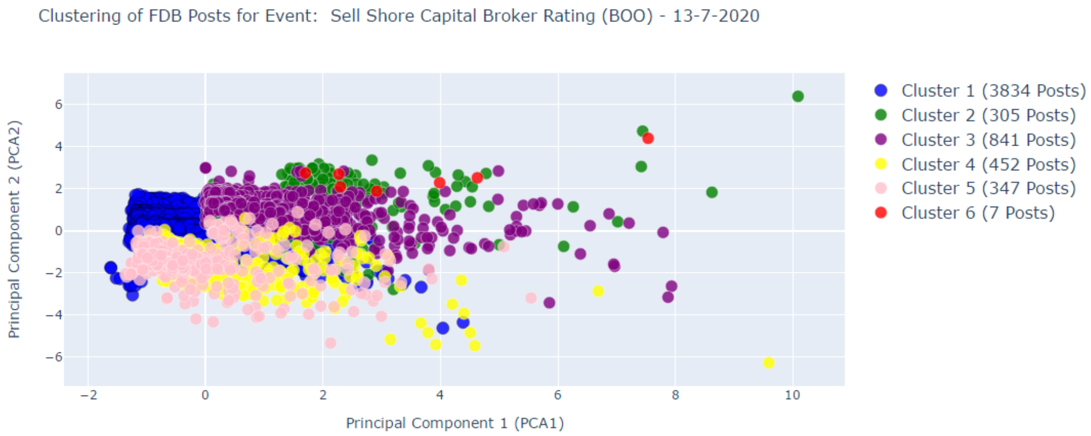
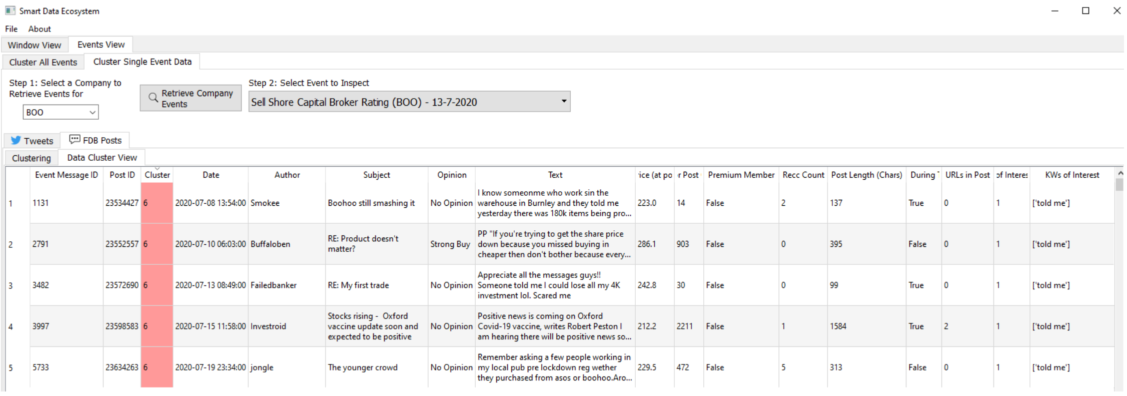


FIGURE 8.7: Clustering of FDB Posts for Boohoo PLC (LON:BOO) for Event: Sell rating by Shore Capital 13/07/2020

8.8 Chapter Summary

This chapter has presented the detection capabilities of the SDE, which take the form of clustering algorithms that are visualised by generating and plotting the two principal components of the respective feature set on a two-dimensional space.



Step 1: Select a Company to Retrieve Events for: BOO

Step 2: Select Event to Inspect: Sell Shore Capital Broker Rating (BOO) - 13-7-2020

	Event Message ID	Post ID	Cluster	Date	Author	Subject	Opinion	Text	ice (at po	r Post	Premium Member	Recc Count	Post Length (Chars)	During	URLs in Post	of Interes	KWs of Interest
1	1131	23534427	6	2020-07-08 13:54:00	Smokee	Boohoo still smashing it	No Opinion	I know someonme who work sin the warehouse in Burnley and they told me yesterday there was 180k items being pro...	223.0	14	False	2	137	True	0	1	['told me']
2	2791	23552557	6	2020-07-10 06:03:00	Buffaloben	RE: Product doesn't matter?	Strong Buy	PP "If you're trying to get the share price down because you missed buying in cheaper then don't bother because every...	286.1	903	False	0	395	False	0	1	['told me']
3	3482	23572690	6	2020-07-13 08:49:00	Failedbanker	RE: My first trade	No Opinion	Appreciate all the messages guys!! Someone told me I could lose all my 4K investment lol. Scared me	242.8	30	False	0	99	True	0	1	['told me']
4	3997	23598583	6	2020-07-15 11:58:00	Investroid	Stocks rising - Oxford vaccine update soon and expected to be positive	No Opinion	Positive news is coming on Oxford Covid-19 vaccine, writes Robert Peston I am hearing there will be positive news so...	212.2	2211	False	1	1584	True	2	1	['told me']
5	5733	23634263	6	2020-07-19 23:34:00	jongle	The younger crowd	No Opinion	Remember asking a few people working in my local pub pre lockdown reg wether they purchased from asos or boohoo.Aro...	229.5	472	False	5	313	False	0	1	['told me']

FIGURE 8.8: FDB posts data points view (Event: Sell rating by Shore Capital)

The clustering process can only commence once non-LSE and cryptocurrency tweets have been successfully filter out by the data layer's cashtag collision methodology. It is therefore prudent to note that without the existence of the cashtag collision classifiers to resolve such conflicts (Section 6.6), filtering of tweets would be a fruitless endeavour, as noisy cryptocurrency tweets would undoubtedly add an additional layer of unnecessary complexity to the clustering process.

Different scenarios that use the clustering approached discussed in this chapter (event-based, tweets, and FDB posts), will be explored further in the next chapter when the SDEs effectiveness is evaluated by conducting qualitative interviews with five financial market experts.

The next chapter will evaluate various tools within the ecosystem, which include resolving cashtag collisions, assessing the credibility of tweets, and the clustering capabilities.

The main contributions of this chapter are as follows:

- Clustering algorithms that operate on LSE tweets that are filtered using the cashtag collision methodology outlined in Section 6.6.
- A front-end interface to visualise the significant clusterings based on the clusters' silhouette scores.
- A data cluster view that allows closer inspection of the data points that belong to each cluster

Chapter 9

Ecosystem Evaluation

9.1 Overview

This chapter reports on an evaluative study undertaken that examines various tools within the Smart Data Ecosystem (SDE) through the use of qualitative interviews with participants that have knowledge of financial markers. The purpose of this chapter is to address the final research question posed in Section 1.5: *Can a smart data ecosystem, through visualisation tools, assist a user in establishing the significance of detected irregularities?*.

The methodology of this evaluation will firstly be introduced (Section 9.2), which includes how participants will be identified and recruited, and will also introduce six scenario that focus on different aspects of the SDE, using specific companies and dates to provide concrete examples to discuss with the participants. Each of the scenarios, including the answers to scenario-specific questions, and the discussion that led from such questions, are presented in Section 9.5.

9.2 Evaluation Methodology

This section will detail how participants were shortlisted (Section 9.2.1), including the inclusion and exclusion criteria, the information given to participants prior to the interviews, and how the interviews were conducted. The hypothesis being tested in this chapter is: An ecosystem that can offer visualisation tools to assist users in establishing if certain data points could be potentially irregular can benefit a regulatory body.

9.2.1 Participant Identification and Shortlisting

To identify suitable participants to take part in this qualitative evaluation, online profiles of staff members belonging the the Department of Accounting, Finance and Banking at Manchester Metropolitan University were visited. For every staff profile, if keywords relating to this research (e.g. stock market, financial market, brokerage) were found, the name would be added to a shortlist. Once the shortlist was complete, the names of suitable candidates were discussed amongst the supervisory team. In total, nine people were shortlisted and emailed inviting them to take part in this study, with five people accepting, two declining, and the remaining two not offering a response. Each of these invitations included a participant information sheet (Appendix I) which gave an overview of the research project as a whole (development of a smart data ecosystem to monitor stock discussion), and what the interview would involve. The invited participants were also informed of the EthOS (MMUs ethics system) number (34325), the ethical approval of this study can be found in Appendix J.

9.2.2 Ecosystem Scenarios

Before the commencement of any interviews with the participants, Six scenarios were designed that focus on a specific company for the purpose of evaluating a specific tool within the SDE. These six scenarios were chosen to showcase the different tools of the SDE and allow for discussions to relate to specific aspects of the SDE. The first two scenarios focused on the SDE's ability to resolve cashtag collisions. The third scenario honed in on the SDE's credibility classifier. The remaining three scenarios looked at the clustering functionality of the SDE. Any prerequisite knowledge required for participants to understand these scenarios will be delivered by presenting a short presentation prior to the scenarios being presented (discussed in Section 9.3). Each of these scenarios will now be outlined, with the discussion that stemmed from these scenarios in the interviews reserved for Section 9.5.

Scenario 1 - Resolving cashtag collisions

The first scenario asked the participants to comment on the SDE's ability to resolve cashtag collisions between two companies listed on different exchanges that possess an identical cashtag.

Participants were shown two sets of tweets originating from the SDE: (1) tweets classified as referencing the LSE variant of the cashtag $TSCO (Tesco PLC - LON:TSCO), and (2) tweets classified as not referencing the LSE (e.g. The Tractor Supply Company - NASDAQ:TSCO). All tweets for this scenario were tweeted and collected on 14/06/2019, in which ten tweets were collected by the SDE containing the $TSCO cashtag. Four of these tweets were classified by the SDE as not belonging to the LSE company (Tesco PLC), with the remaining six being classified as belonging to Tesco PLC. Participants are shown the tweets, including the classification assigned by the SDE, and asked to comment if they agree with SDE classification of the tweets.

Participants will be asked the following questions as part of this scenario:

1. Do you agree that the ecosystem has been successful in distinguishing between tweets referring to Tesco and tweets referring to the Tractor Supply Company?
2. How helpful do you think this functionality is for investors and automated tools?

Scenario 2 - Filtering out cryptocurrency tweets

The second scenario focuses on the SDE's ability to filter out noisy cryptocurrency tweets. Participants were shown tweets containing the $NANO cashtag. This cashtag refers to Nanoco Group PLC on the LSE and also refers to the popular cryptocurrency, Nano.

In total, 112 tweets for a 1-day period were retrieved for 03/07/19 containing the $NANO cashtag. One tweet was classified as relating to the LSE company, with the remaining 111 tweets classified as not relating to the LSE (non-LSE exchange or cryptocurrency tweets). Participants will be given the opportunity to browse all of the tweets containing the $NANO cashtag for this time period.

Participants will be asked one question as part of this scenario: *Do you agree that the ecosystem can successfully navigate noisy cryptocurrency tweets and only highlight tweets referring to the LSE-listed company?*

Scenario 3 - Assessing credibility of financial stock tweets

The third scenario focused on the credibility class assigned to tweets by the credibility classifier discussed in Chapter 8. This is arguably the most subjective part of this research due to the subjective nature of assessing credibility as discussed in Section 7.2. Participants will be shown tweets from each of the credibility labels (not credible, ambiguous, and credible), and discuss what they believe makes a financial tweet credible. Any reference to tweets from this point forward relates to tweets classified as relating to an LSE company.

Scenario 4 - Clustering of events

The fourth scenario introduces the clustering capabilities of the ecosystem, utilising the methodology discussed in Chapter 8. Events for the company AstraZeneca (LON:AZN) were clustered into two groups. AstraZeneca events were chosen primarily due to prominence of the company at that time due to the COVID vaccine news, allowing deeper discussion regarding the events surrounding the company within the interviews. The intention of this k value is to group the events into two distinct groupings - *regular* and *irregular*. Once the events for the company were clustered, participants were asked: *"Based on the clustering performed, which cluster(s) would be of particular interest for further investigation?"*

Scenario 5 - Clustering of tweets within an event

The fifth scenario involves evaluating the ecosystem's ability to cluster tweets within an event. A single event was chosen: Shore Capital's buy rating for AstraZeneca (LON:AZN) on 14/07/2020. This event was chosen as the silhouette score for the clustering of tweets for this event shows that three clusters is the most optimal, which should prompt for a more interest discussion versus an event with only a few tweets in two clusters.

Scenario 6 - Clustering of FDB posts within an event

The final scenario involves evaluating the ecosystem's ability to cluster FDB posts within an event. This scenario focused on an event for GlaxoSmithKline PLC (LON:GSK) in which the broker, Berenberg Bank, issued a buy rating on 28/09/2020. Participants were initially asked the question: *"Based on the clustering of FDB posts for this event, which data points / cluster appear to be anomalous?"*

9.3 Interview Overview

The interview portion of evaluation methodology (Figure 9.1) depicts how each of the interviews was conducted. All of the interviews were conducted virtually over Microsoft Teams. At the start of the interview, participants were shown a ten-minute presentation (Appendix H). This presentation included a high-level description of the SDE, the data sources used by the SDE, and the issue of cashtag collisions. Events were also introduced (as several of the scenarios focus on specific company events) followed by an explanation of the clustering functionality used by the SDE.

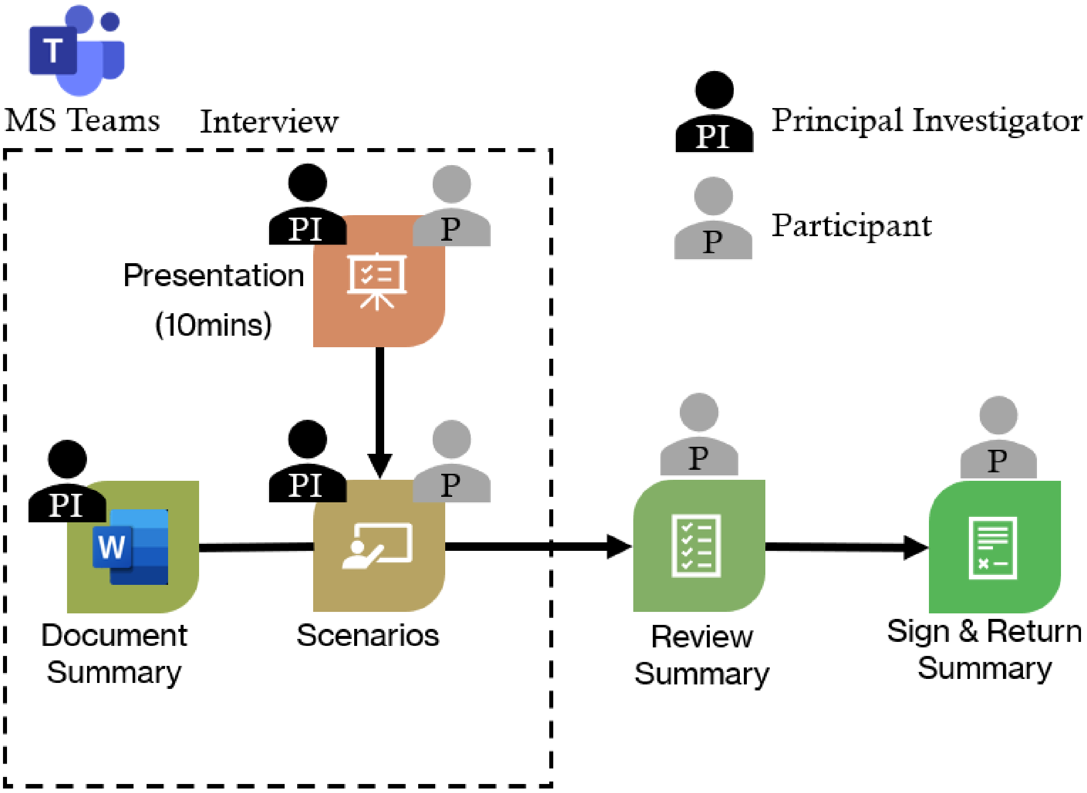


FIGURE 9.1: Evaluation methodology (interview stage)

After the presentation, each participant was briefly introduced to the SDE GUI. The participant was then shown six scenarios (Section 9.5), each of which focused on specific functionality within the SDE. Each of these scenarios used a specific company and time/event window which was consistent across all participants. During each of these scenarios, the Principal Investigator (PI) documented the answers and discussion following the questions. After the six scenarios had been discussed, closing questions were asked to ascertain the participant's overall thoughts on the SDE. At the end of the interview, the PI sent the documented summary to the participant to review and sign-off if they were satisfied with the summary. In the event the participant wanted to amend any information, they were invited to enable 'tracked changes' within Microsoft Word to amend or add any supplementary information to the document and then send it back to the PI.

An overview of the five participants who agreed to take part in this evaluative study will now be presented.

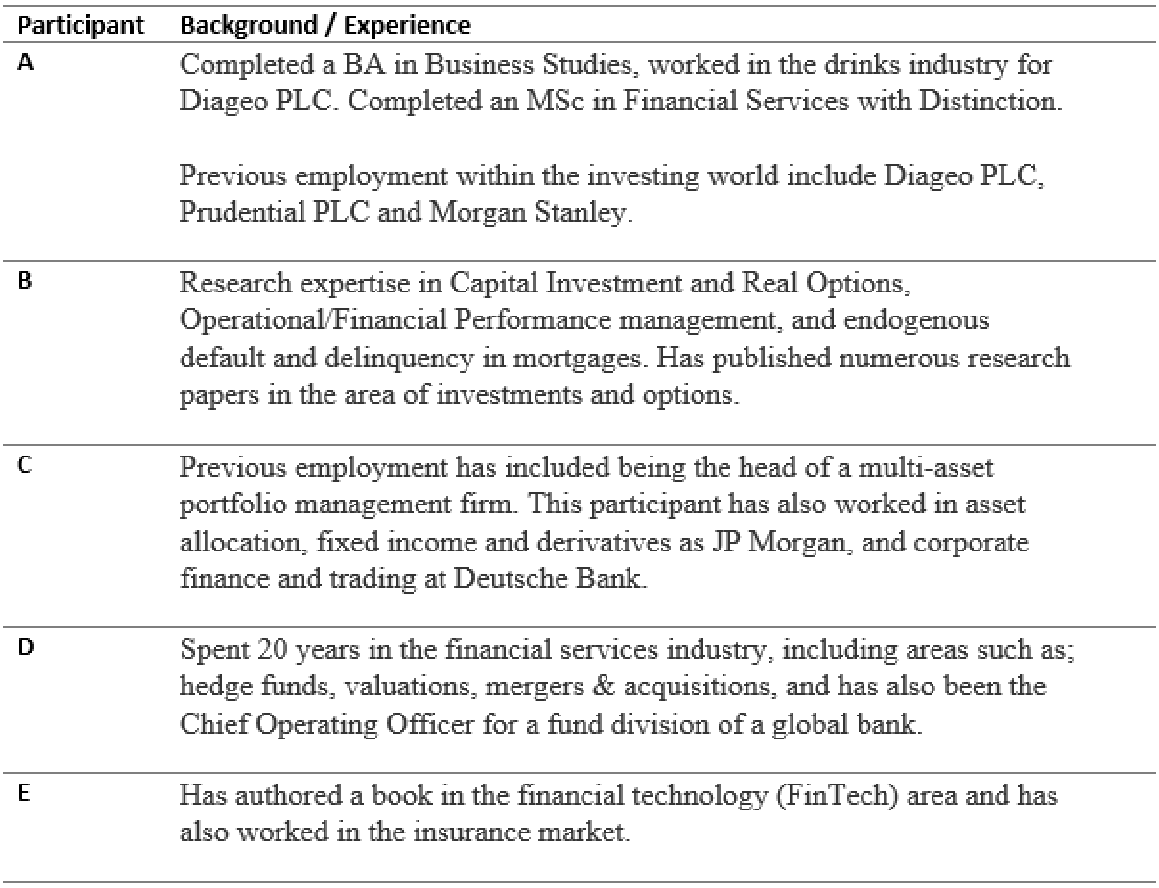
9.4 Participant Overview

An overview of the participants is provided in (Table 9.2). Five participants were chosen, as Alroobaea and Mayhew (2014) found that a sample size of 5-9 participants was enough to highlight around 80% of issues when undertaking a heuristic evaluation.

All participants are currently employed within the Department of Accounting, Finance and Banking at Manchester Metropolitan University. These participants each have knowledge of how financial markets operate, with some (Participants D & E) having an in-depth understanding of the cryptocurrency market.

9.5 Evaluation Scenarios

This subsection will summarise the results of interviewing the five participants shortlisted in Section 9.4, using the scenarios outlined in Section 9.2.2.



Participant	Background / Experience
A	Completed a BA in Business Studies, worked in the drinks industry for Diageo PLC. Completed an MSc in Financial Services with Distinction. Previous employment within the investing world include Diageo PLC, Prudential PLC and Morgan Stanley.
B	Research expertise in Capital Investment and Real Options, Operational/Financial Performance management, and endogenous default and delinquency in mortgages. Has published numerous research papers in the area of investments and options.
C	Previous employment has included being the head of a multi-asset portfolio management firm. This participant has also worked in asset allocation, fixed income and derivatives as JP Morgan, and corporate finance and trading at Deutsche Bank.
D	Spent 20 years in the financial services industry, including areas such as; hedge funds, valuations, mergers & acquisitions, and has also been the Chief Operating Officer for a fund division of a global bank.
E	Has authored a book in the financial technology (FinTech) area and has also worked in the insurance market.

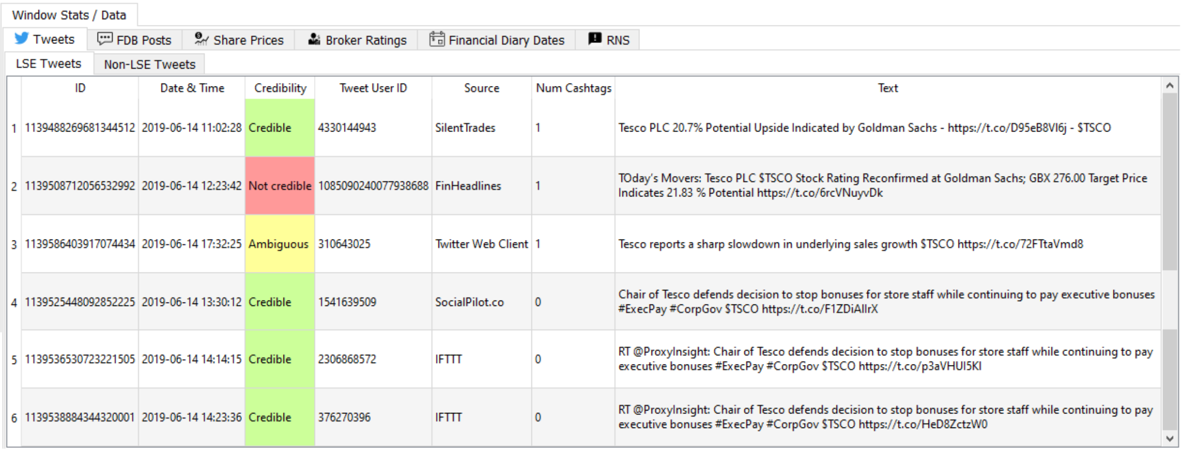
FIGURE 9.2: Participant Overview

9.5.1 Scenario 1 - Resolving Cashtag Collisions between Companies

Participants were initially asked: *"Do you agree that the ecosystem has been successful in distinguishing between tweets referring to Tesco and tweets referring to the Tractor Supply Company?"*. All participants agreed that the six LSE tweets for this time window (Figure 9.3) all featured the name of the LSE company and were undoubtedly referring to the LSE company.

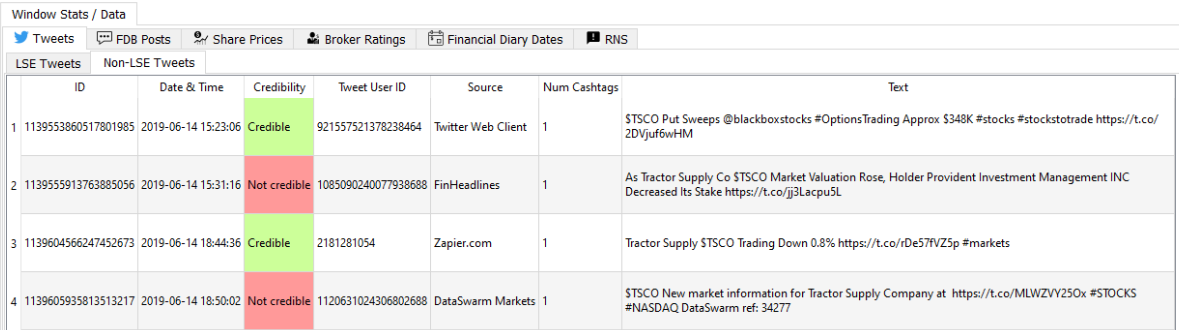
Participants were then shown the tweets classified as not relating to the LSE company (Figure 9.4). Four non-LSE tweets featured in this time-slice window. Tweets 2-4 all feature the Tractor Supply Company name, and all participants agreed that the presence of the name was enough to resolve the cashtag collision conflict. However, the first non-LSE tweet (Tweet #1, Figure 9.4) does not mention either company by name. Participants were asked to identify any other tweet characteristics that could help determine which company the tweet was referencing.

Participant A indicated that the presence of a dollar symbol could be enough



	ID	Date & Time	Credibility	Tweet User ID	Source	Num Cashtags	Text
1	1139488269681344512	2019-06-14 11:02:28	Credible	4330144943	SilentTrades	1	Tesco PLC 20.7% Potential Upside Indicated by Goldman Sachs - https://t.co/D95eB8Vl6j - $TSCO
2	1139508712056532992	2019-06-14 12:23:42	Not credible	1085090240077938688	FinHeadlines	1	TOday's Movers: Tesco PLC $TSCO Stock Rating Reconfirmed at Goldman Sachs; GBX 276.00 Target Price Indicates 21.83 % Potential https://t.co/6rcVNuyvDk
3	1139586403917074434	2019-06-14 17:32:25	Ambiguous	310643025	Twitter Web Client	1	Tesco reports a sharp slowdown in underlying sales growth $TSCO https://t.co/72FTtaVmd8
4	1139525448092852225	2019-06-14 13:30:12	Credible	1541639509	SocialPilot.co	0	Chair of Tesco defends decision to stop bonuses for store staff while continuing to pay executive bonuses #ExecPay #CorpGov $TSCO https://t.co/F1ZDiAllrX
5	1139536530723221505	2019-06-14 14:14:15	Credible	2306868572	IFTTT	0	RT @ProxyInsight: Chair of Tesco defends decision to stop bonuses for store staff while continuing to pay executive bonuses #ExecPay #CorpGov $TSCO https://t.co/p3aVHUI5KI
6	1139538884344320001	2019-06-14 14:23:36	Credible	376270396	IFTTT	0	RT @ProxyInsight: Chair of Tesco defends decision to stop bonuses for store staff while continuing to pay executive bonuses #ExecPay #CorpGov $TSCO https://t.co/HeD8ZctzW0

FIGURE 9.3: Scenario 1 LSE Tweets



	ID	Date & Time	Credibility	Tweet User ID	Source	Num Cashtags	Text
1	1139553860517801985	2019-06-14 15:23:06	Credible	921557521378238464	Twitter Web Client	1	$TSCO Put Sweeps @blackboxstocks #OptionsTrading Approx $348K #stocks #stockstotrade https://t.co/2DVjuf6wHM
2	1139555913763885056	2019-06-14 15:31:16	Not credible	1085090240077938688	FinHeadlines	1	As Tractor Supply Co $TSCO Market Valuation Rose, Holder Provident Investment Management INC Decreased Its Stake https://t.co/jj3Lacpu5L
3	1139604566247452673	2019-06-14 18:44:36	Credible	2181281054	Zapier.com	1	Tractor Supply $TSCO Trading Down 0.8% https://t.co/rDe57fVZ5p #markets
4	1139605935813513217	2019-06-14 18:50:02	Not credible	1120631024306802688	DataSwarm Markets	1	$TSCO New market information for Tractor Supply Company at https://t.co/MLWZVY25Ox #STOCKS #NASDAQ DataSwarm ref: 34277

FIGURE 9.4: Scenario 1 Non-LSE Tweets

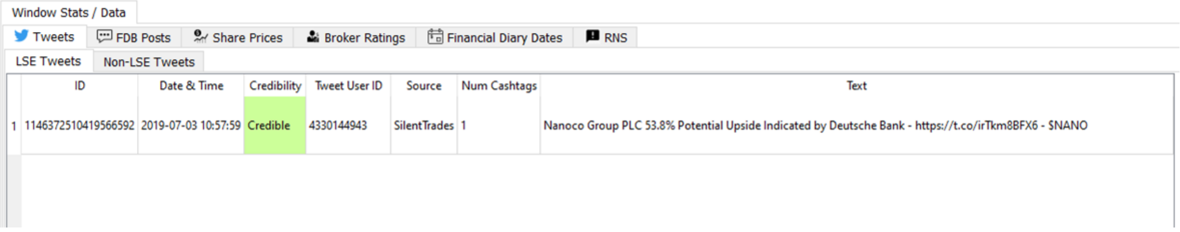
information to determine the tweet references a non-LSE stock. However, they did add that some investors will use the currency found in the company report instead of the exchange the company is listed on when communicating about the stock. Participant A agreed that the tweet was classified correctly. Several participants (B, C, E) noted that some companies have share quotations on other stock exchanges – meaning investors and companies may use different currency symbols when referencing a company. This indicates that the currency will not always be a perfect predictor in establishing the stock exchange being referred to in a tweet.

Scenario Summary

The presence of the LSE company name in the tweets largely lead participants to be confident that the classifier was successful in resolving the colliding cashtag tweets. However, it is important to note that relying on that approach will only be effective if the companies sharing the cashtag are separate companies, and therefore have different company names. Several companies (e.g. Vodafone PLC) have listings on multiple exchanges, and will therefore have the same name and CEO, meaning relying solely upon the name to resolve this conflict would not be a reliable approach.

9.5.2 Scenario 2 - Filtering Noisy Cryptocurrency Tweets

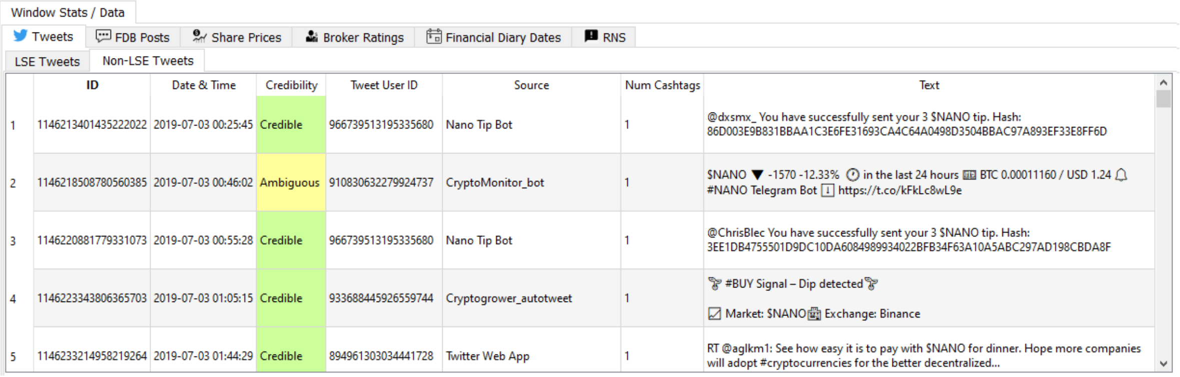
To begin with, participants were shown the tweet classified as referencing an LSE company (Figure 9.5).



	ID	Date & Time	Credibility	Tweet User ID	Source	Num Cashtags	Text
1	1146372510419566592	2019-07-03 10:57:59	Credible	4330144943	SilentTrades	1	Nanoco Group PLC 53.8% Potential Upside Indicated by Deutsche Bank - https://t.co/irTkm8BFX6 - $NANO

FIGURE 9.5: Scenario 2 LSE Tweets

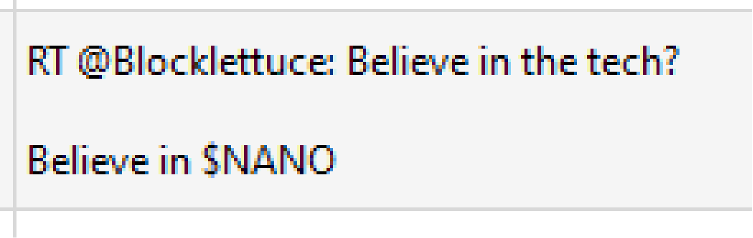
All participants agreed that this tweet references the LSE-listed stock due to the presence of the LSE company name. Participants were then shown the tweets classified as not relating to the LSE (Figure 9.6). Participants (C, D, E) noted that various terms found within the tweets' text lent support to the tweets related to cryptocurrencies. Specifically, terms including: ″binance″, ″mining″, ″crypto″, ″coin″, and ″crypto wallet″ are associated with cryptocurrency trading.



	ID	Date & Time	Credibility	Tweet User ID	Source	Num Cashtags	Text
1	1146213401435222022	2019-07-03 00:25:45	Credible	966739513195335680	Nano Tip Bot	1	@dxsmx_ You have successfully sent your 3 $NANO tip. Hash: 86D003E9B831BBAA1C3E6FE31693CA4C64A0498D3504BBAC97A893EF33E8FF6D
2	1146218508780560385	2019-07-03 00:46:02	Ambiguous	910830632279924737	CryptoMonitor_bot	1	$NANO ▼ -1570 -12.33% 🕐 in the last 24 hours 💵 BTC 0.00011160 / USD 1.24 🔔 #NANO Telegram Bot ⬇ https://t.co/kFkLc8wL9e
3	1146220881779331073	2019-07-03 00:55:28	Credible	966739513195335680	Nano Tip Bot	1	@ChrisBlec You have successfully sent your 3 $NANO tip. Hash: 3EE1DB4755501D9DC10DA6084989934022BFB34F63A10A5ABC297AD198CBDA8F
4	1146223343806365703	2019-07-03 01:05:15	Credible	933688445926559744	Cryptogrower_autotweet	1	🚁 #BUY Signal – Dip detected 🚁 📈 Market: $NANO 🏦 Exchange: Binance
5	1146233214958219264	2019-07-03 01:44:29	Credible	894961303034441728	Twitter Web App	1	RT @aglkm1: See how easy it is to pay with $NANO for dinner. Hope more companies will adopt #cryptocurrencies for the better decentralized...

FIGURE 9.6: Scenario 2 Non-LSE Tweets

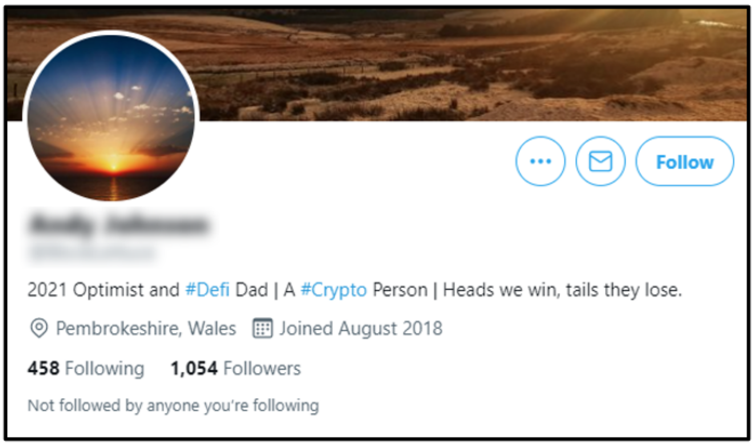
Participant A noted that the textual content will often not be enough for an investor to determine if a tweet references a stock or a cryptocurrency tweet. The tweet in Figure 9.7, for example, is short in length and does not reference any particular company or cryptocurrency by name. We did, however, trace the tweet ID to the user profile who tweeted it (Figure 9.8) and the participant agreed that the biography of the user indicated the user was a cryptocurrency trader.



RT @Blocklettuce: Believe in the tech?

Believe in $NANO

FIGURE 9.7: Scenario 2 Ambiguous Tweet Text



[redacted]

2021 Optimist and #Defi Dad | A #Crypto Person | Heads we win, tails they lose.

Pembrokeshire, Wales Joined August 2018

458 Following 1,054 Followers

Not followed by anyone you're following

FIGURE 9.8: Scenario 2 Ambiguous Tweet User Profile

Participant D highlighted that the tone used within tweets is also helpful – certain phrases such as ”buy your steak dinner with NANO” are less formal - and would associate such tweets with cryptocurrency trading. Participant E, an expert in the cryptocurrency space, also noted that cryptocurrency tweets tend to be more speculative and typically include terms that emphasise the urgency to invest (e.g. ”HUGE GAINS - SIGN UP NOW!”). This participant also mentioned that the tweet's source is also an important characteristic to consider, as many cryptocurrency tweets originate from automated (bot) accounts.

Scenario Summary

The use of informal 'slang' and promises of huge returns in cryptocurrencies are certainly one of the most distinguishable hallmarks for identifying cryptocurrency tweets. Other characteristics which can help in identify cryptocurrency tweets are a large presence of emojis that indicate the cryptocurrency price is "taking off" - the rocket emoji being commonly used to illustrate this.

In respect to the first two scenarios, four participants (A, C, D, E) agreed that filtering such tweets would be of benefit to investors who were not cryptocurrency traders. The same four participants agreed that this filtering aspect would be a critical pre-processing step from a regulatory standpoint of monitoring the discussion of stocks on Twitter.

9.5.3 Scenario 3 - Financial Stock Tweet Credibility

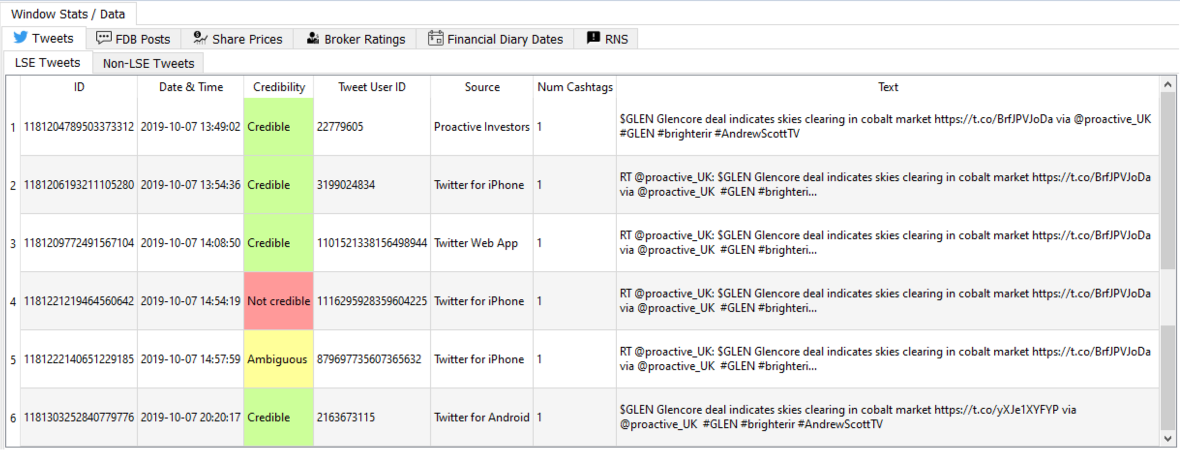
Participants were given a brief explanation of the credibility classifier utilised by the SDE. This included a list of features that the classifier is trained on, based on our work in Evans, Owda, Crockett, and Ana Fernandez Vilas (2021) - which found the account age of the user and their number of followers to be two of the most informative features in establishing credibility.

Participants were shown tweets relating to Glencore PLC (LON:GLEN). These tweets (Figure 9.9) contain near-identical textual content but have been tweeted by different users. As the credibility classifier adopted by the SDE utilities over thirty features, participants were shown the tweet text, along with the two most informative features (according to Evans, Owda, Crockett, and Ana Fernandez Vilas (ibid.)) for assessing credibility - user account age, and the number of followers. The text, along with the two features, were shown in a table format within the presentation to aid the users in comparing the tweets. The participants were also free to inspect other tweet features during the scenario discussion.

Participants were initially asked the following question: *"Do you agree with the ecosystem's decision to mark the selected tweet as not credible? I.e. Would you use the features (e.g. age of the account, number of followers) as a measure of credibility?"*

Participant A would be swayed by those characteristics of the tweet but noted that it is important to verify they have the background knowledge associated with any information/claims in the tweet.

Participant B stated he would not trust a user who has been active for ten years more than a user who had only had an account for three months. This participant also noted that it would be interesting to see the breakdown of a user's followers



	ID	Date & Time	Credibility	Tweet User ID	Source	Num Cashtags	Text
1	1181204789503373312	2019-10-07 13:49:02	Credible	22779605	Proactive Investors	1	$GLEN Glencore deal indicates skies clearing in cobalt market https://t.co/BrfJPVJoDa via @proactive_UK #GLEN #brighterir #AndrewScottTV
2	1181206193211105280	2019-10-07 13:54:36	Credible	3199024834	Twitter for iPhone	1	RT @proactive_UK: $GLEN Glencore deal indicates skies clearing in cobalt market https://t.co/BrfJPVJoDa via @proactive_UK #GLEN #brighteri...
3	1181209772491567104	2019-10-07 14:08:50	Credible	1101521338156498944	Twitter Web App	1	RT @proactive_UK: $GLEN Glencore deal indicates skies clearing in cobalt market https://t.co/BrfJPVJoDa via @proactive_UK #GLEN #brighteri...
4	1181221219464560642	2019-10-07 14:54:19	Not credible	1116295928359604225	Twitter for iPhone	1	RT @proactive_UK: $GLEN Glencore deal indicates skies clearing in cobalt market https://t.co/BrfJPVJoDa via @proactive_UK #GLEN #brighteri...
5	1181222140651229185	2019-10-07 14:57:59	Ambiguous	879697735607365632	Twitter for iPhone	1	RT @proactive_UK: $GLEN Glencore deal indicates skies clearing in cobalt market https://t.co/BrfJPVJoDa via @proactive_UK #GLEN #brighteri...
6	1181303252840779776	2019-10-07 20:20:17	Credible	2163673115	Twitter for Android	1	$GLEN Glencore deal indicates skies clearing in cobalt market https://t.co/yXJe1XYFYP via @proactive_UK #GLEN #brighterir #AndrewScottTV

FIGURE 9.9: Scenario 3 LSE Tweets

over time - one user may have had all of their followers follow them in the first year and had no new followers since - indicating they may have bought their followers or had become less active. This participant stated that it was not clear what the breakpoints were (e.g. number of followers) in terms of how the classifier assigned the credibility scores.

Participant C noted that the classifier was right to classify the non-credible tweet as not being credible, as a new user to a platform is justified as warranting further speculation. This participant also raised an issue with depending on features such as the account age for assigning credibility. Namely, an experienced investor may sign-up to the platform - and hence have a low account age - meaning the classifier would unfairly associate that user, and the user's tweets, as not being credible. Naturally, other features within such as user's tweet which the classifier also takes into consideration (e.g. count of credible URLs in the tweet) may alleviate such concerns.

Participant D believed features such as the user account age and number of followers are indeed useful features for considering credibility. Naturally, automated (bot) accounts on Twitter do no survive for very long, as flooding Twitter with automated, non-informative tweets is against Twitter's terms of service. Automated bot accounts, therefore, will naturally have an account age that is low with a small number of followers, leading to such tweets being assigned as not credible.

Participant E highlighted that as many of the features are derived from the user (e.g. account age, number of followers/following), the score could be assigned to a user, instead of the tweet. This could lead to the SDE assigning scores to users, and not the tweet themselves. The participant also noted that some tweets were fairly close in terms of account age (5.85 years vs 4.39 years) but were dramatically different in terms of the number of followers (134 vs 1150). This difference could indicate the account which is slightly older with fewer followers is perhaps not an active user, and the ratio of the account age and the number of followers may be an interesting avenue to explore.

Participants were then asked a follow-up question: *"What, in your opinion, makes a tweet credible?"*

Participant A noted that a tweet containing a hyperlink would be a likely indicator of credibility - but only if the hyperlink is related to the content of the tweet. Looking at the user's profile of the tweet and seeing if they have specific qualifications such as being CFA (Chartered Financial Analyst) certified. The participant went on to note that even if a user did list themselves as being possess the CFA certification in their Twitter biography, that alone would not prove the user holds such a certification, and additional research would need to be undertaken to verify such claims.

Participant B shared similar concerns, in that any hyperlinks within a tweet should be subject to additional scrutiny, such as are the hyperlinks associated with a well-known tipster or reputable news company? This participant noted that they only trusted certain platforms (e.g. Bloomberg) that have a proven track record - and that Twitter is not such a platform.

Participant C stated they would not trust Twitter at all, as the market is driven by fear and greed. Also, as tweets are limited by the number of characters they can contain, they are likely to not contain enough information to instil trust in investors, particularly if they do not contain hyperlinks. The tweets themselves would not swing the participant to take action, but would more likely act as a signpost for information.

Participant D raised the importance of the tone of the language within a tweet. Fewer abbreviations and emojis would seem to be more credible than tweets filled with both - rocket emojis and smiley faces do not infer a high amount of credibility. The presence of credible URLs is a good source of establishing the credibility too. Linking to other sources such as Reddit may not be so credible – but that could be a person linking to a legitimate news story involving Reddit, so context is important.

Finally, Participant E noted that the volume of tweets from a user could be indicative of credibility. A user who tweets continuously at a high volume may appear less credible than a user who tweets less frequently with more informative content.

Scenario Summary

Indeed, the credibility aspect of tweets is a contentious issue which has provided different viewpoints from the participants. As the credibility classifiers trained and discussed in Chapter 7 are trained on up to sixty different features, all of which have varying levels of informative power, it becomes increasing difficult to interpret how the classifier reaches its classification decision. Participant A raised the concern that even if you could trace the Twitter user's profile to a more professional account (e.g. LinkedIn), there is no guarantee that their LinkedIn proves they are a reputable investor - as everyone has the capability to lie online.

9.5.4 Scenario 4 - Clustering of Events

The fourth scenario begins with the clustering of events for AstraZeneca PLC (Figure 9.10).

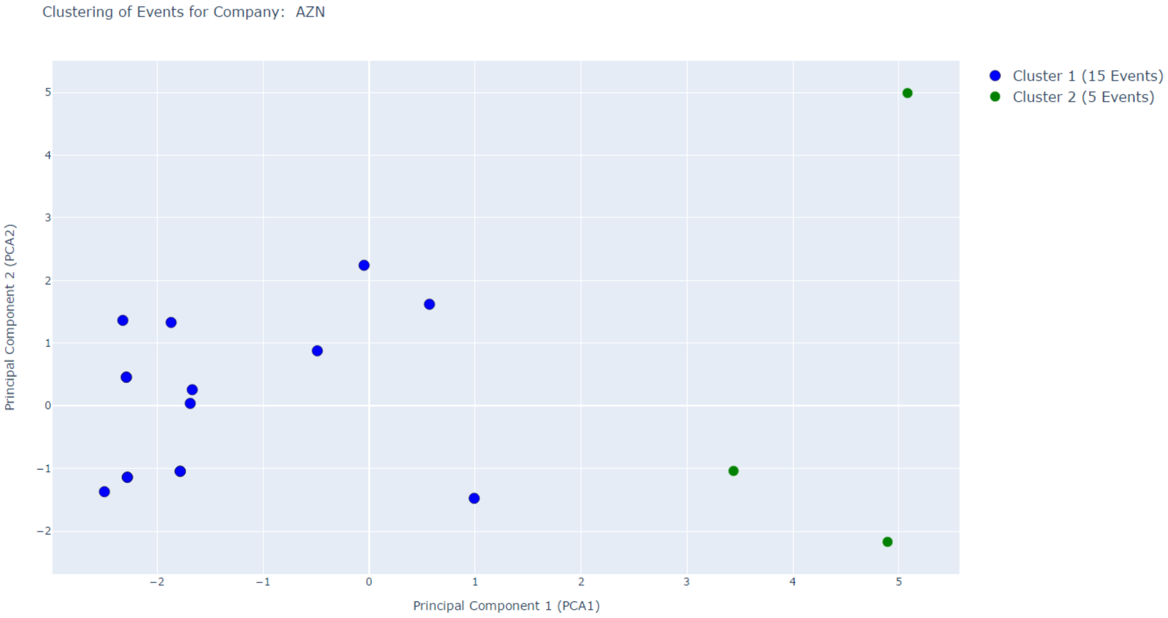
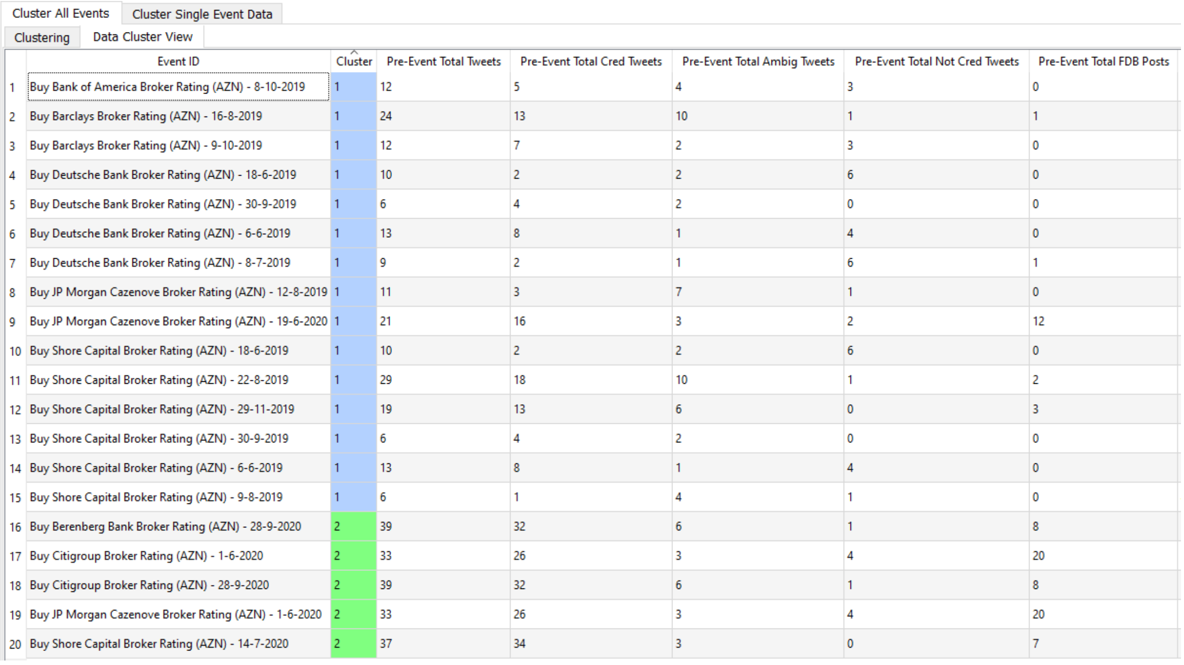


FIGURE 9.10: Scenario 4 - Clustering of AstraZeneca (LON:AZN) Events

Participants A, C & E stated they would initially be interested in the green cluster (outliers). Participant B could not give a definitive answer as to which cluster would be of initial interest, as the graph only shows the clustering output, and not what the clusters themselves indicate (e.g. insider information). Participant C observed that the outliers could be total noise or provide interesting insights. Participant D noted that both clusters should be taken into account, but that it would depend on what their analysis viewpoint was. As an investor, the participant cited they would be interested in cluster 1 where most of the data points are, whereas a regulatory would more likely be focused on the cluster with the least data points to see if irregular activity may be taking place.

The second question posed to the participants was: *Based on viewing the data for the cluster(s) you have selected, do you understand why the data point(s) have been clustered in such a way?*

Participants were shown the "Data Cluster View" tab (Figure 9.11), that shows each of the data points along with columns corresponding to each of the features used to cluster the data points.



	Event ID	Cluster	Pre-Event Total Tweets	Pre-Event Total Cred Tweets	Pre-Event Total Ambig Tweets	Pre-Event Total Not Cred Tweets	Pre-Event Total FDB Posts
1	Buy Bank of America Broker Rating (AZN) - 8-10-2019	1	12	5	4	3	0
2	Buy Barclays Broker Rating (AZN) - 16-8-2019	1	24	13	10	1	1
3	Buy Barclays Broker Rating (AZN) - 9-10-2019	1	12	7	2	3	0
4	Buy Deutsche Bank Broker Rating (AZN) - 18-6-2019	1	10	2	2	6	0
5	Buy Deutsche Bank Broker Rating (AZN) - 30-9-2019	1	6	4	2	0	0
6	Buy Deutsche Bank Broker Rating (AZN) - 6-6-2019	1	13	8	1	4	0
7	Buy Deutsche Bank Broker Rating (AZN) - 8-7-2019	1	9	2	1	6	1
8	Buy JP Morgan Cazenove Broker Rating (AZN) - 12-8-2019	1	11	3	7	1	0
9	Buy JP Morgan Cazenove Broker Rating (AZN) - 19-6-2020	1	21	16	3	2	12
10	Buy Shore Capital Broker Rating (AZN) - 18-6-2019	1	10	2	2	6	0
11	Buy Shore Capital Broker Rating (AZN) - 22-8-2019	1	29	18	10	1	2
12	Buy Shore Capital Broker Rating (AZN) - 29-11-2019	1	19	13	6	0	3
13	Buy Shore Capital Broker Rating (AZN) - 30-9-2019	1	6	4	2	0	0
14	Buy Shore Capital Broker Rating (AZN) - 6-6-2019	1	13	8	1	4	0
15	Buy Shore Capital Broker Rating (AZN) - 9-8-2019	1	6	1	4	1	0
16	Buy Berenberg Bank Broker Rating (AZN) - 28-9-2020	2	39	32	6	1	8
17	Buy Citigroup Broker Rating (AZN) - 1-6-2020	2	33	26	3	4	20
18	Buy Citigroup Broker Rating (AZN) - 28-9-2020	2	39	32	6	1	8
19	Buy JP Morgan Cazenove Broker Rating (AZN) - 1-6-2020	2	33	26	3	4	20
20	Buy Shore Capital Broker Rating (AZN) - 14-7-2020	2	37	34	3	0	7

FIGURE 9.11: Scenario 4 - AstraZeneca (LON:AZN) Clustering Data Points

Participants A & C stated that it was not immediately clear why the data points had been clustered in the way they had been. Participant A noted that the post-event tweet seemed higher (for the events in the minority green cluster). This participant also noted how pre-event tweets is a good indicator of potential insider trading, as some investors may be discussing news that may not yet be made public and, as a result, a broker may yet to provide a rating. Participant C noted it was not clear what the rules were in deciding how the data points are clustered. They did note it appeared there was more activity for the events in cluster 2, but it's not clear if that is the deciding factor in determining how the data points are assigned to a cluster. They added that it would take a lot of time for them to actually look at all of the values for the different data points in each cluster.

Participant B said it was not immediately clear why the data points in cluster 2 were assigned that cluster. They did, however, mention that the stock market now is at a different level that it was a year ago - almost all of the data points in cluster 1 occur in 2019, with all data points in cluster 2 occurring in 2020. This participant cited they would be suspicious of the clustering approach for this reason.

Participants D & E highlighted that the green cluster had higher values for the pre-event window - the number if unique Twitter users in that window is higher - indicating more users are involved in the discussion. Participant D added additional feedback to aid the interpretation of the clustering, such as providing statistical measures of all of the data points in each cluster to get a summary of the different clusters (e.g. average pre-event unique Twitter users in Cluster X).

Scenario Summary

The clustering of events for a company serves as a high-level clustering approach, in which the events themselves are clustered, which can serve as an initial 'screening'. Clustering of tweets and FDB posts can then follow, either on the events clustered as being *normal* or *irregular*. The feedback relating to summarising the data points via statistical measures would be invaluable in terms of speeding up the interpretation process and allowing quicker insights to be made regarding the clustering process.

9.5.5 Scenario 5 - Clustering of Event Tweets

Figure 9.12 shows the clustering output for tweets within this event window.

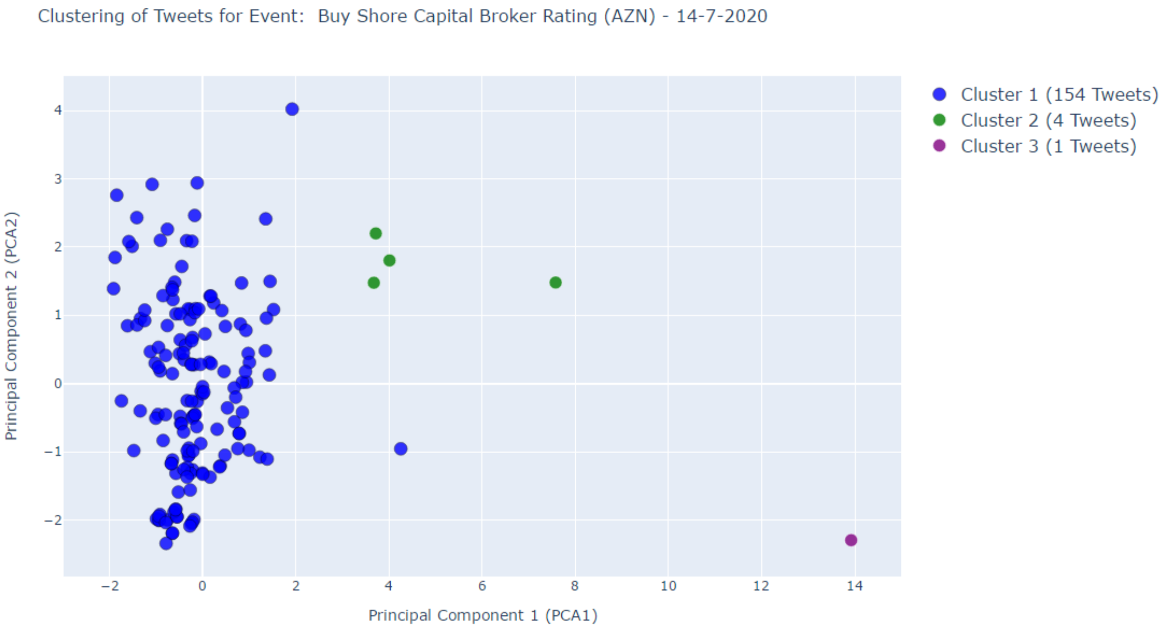


FIGURE 9.12: Scenario 5 - Clustering of Tweets for Shore Capital Buy Broker Rating - AstraZeneca (LON:AZN) Event 14/7/2020

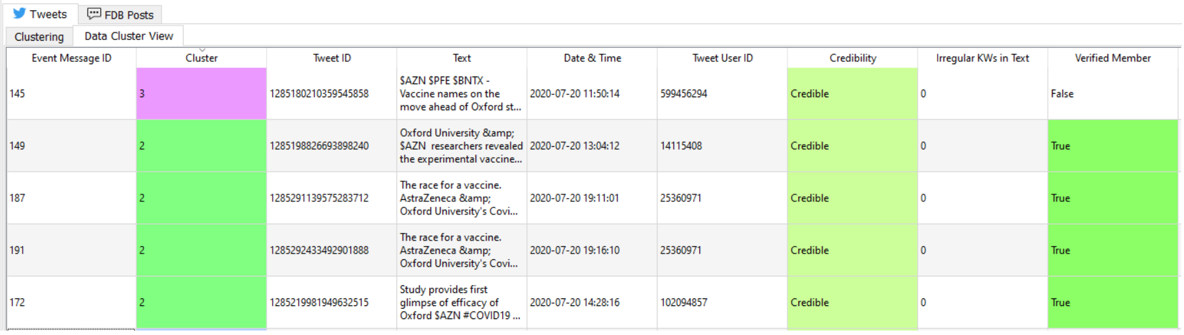
Participants were initially asked *Based on the clustering of tweets for this event, which data point(s) / cluster(s) appear to be anomalous?*

Participants A, C, D and E stated they would focus initially on the outliers to ascertain why they have been clustered differently. Participant D added that as an investor, they might be interested in cluster 1, where most of the data points lie.

Participant B noted that it was not clear what the graph was showing (e.g. was it showing insider information)? They cited that the clustering process only displaying number, and not meaningful labels was making the initial analysis difficult to undertake.

Participants were then asked: ”Based on viewing the data point(s) for this cluster(s), is it clear why the ecosystem has decided to cluster the data point(s) in this way?”. Participants were shown the individual data points, along with the features used to perform the clustering (Figure 9.13).

Participant A initially noted that the word ”Oxford” appears to be in the tweet text for tweets in cluster 2, they went on to clarify that the four data points belonging to cluster 2 were all made by verified members (Participants D & E made the same observation). This participant also shared concerns of verified members circulating information on stocks - non-expert investors may be swayed by verified accounts



Event Message ID	Cluster	Tweet ID	Text	Date & Time	Tweet User ID	Credibility	Irregular KWs in Text	Verified Member
145	3	1285180210359545858	$AZN $PFE $BNTX - Vaccine names on the move ahead of Oxford st...	2020-07-20 11:50:14	599456294	Credible	0	False
149	2	1285198826693898240	Oxford University &amp; $AZN researchers revealed the experimental vaccine...	2020-07-20 13:04:12	14115408	Credible	0	True
187	2	1285291139575283712	The race for a vaccine. AstraZeneca &amp; Oxford University's Covi...	2020-07-20 19:11:01	25360971	Credible	0	True
191	2	1285292433492901888	The race for a vaccine. AstraZeneca &amp; Oxford University's Covi...	2020-07-20 19:16:10	25360971	Credible	0	True
172	2	1285219981949632515	Study provides first glimpse of efficacy of Oxford $AZN #COVID19 ...	2020-07-20 14:28:16	102094857	Credible	0	True

FIGURE 9.13: Scenario 5 - Data points View

sharing stock information - and that verified members (e.g. celebrities) may have no knowledge of stocks.

Participant B & D noted that the only data point in cluster 3 was created by a user with a huge number of following count (i.e. the user follows many other accounts). Participant B also stated that from a regulatory standpoint, it would be useful if clusters were labelled to highlight what the clusters contained (e.g. Cluster X contains tweets suspected of containing of insider information). Participant D added that the data points within cluster 3 also appear to have a higher count of URLs and hashtags compared to data points in the other clusters - indicating they may have more information contained within them.

Participant C stated it would take them a long time to analyse the data and to come to a conclusion as to why the data points were clustered in such a way.

Participants D & E added that, as with the previous scenario, summarising the clusters by providing statistical measures of the data points within each cluster would assist in the interpretation of the clusters.

Scenario Summary

The clustering of tweets undoubtedly provides an immediate viewpoint into the different types of messages circulating on Twitter. As tweets made by verified members are very rare in the event windows within the SDE, verified user tweets should warrant special attention. Verified members typically have large follower bases, which could be interpreted as a means of credibility. It is important that such users are not allowed to spread misinformation on stocks, as they will likely have a bigger impact on the share price. As participant B stated, it is not immediately clear what any clustering graph shows in respect to *why* data points have been clustered the way they have been. For that reason, it is important to know that unsupervised approaches such as clustering will only give you the clusters, and not the rules associated with how the clusters were formed.

9.5.6 Scenario 6 - Clustering of Event FDB Posts

The final scenario focused on the clustering of an individual event's FDB posts (Figure 9.14).

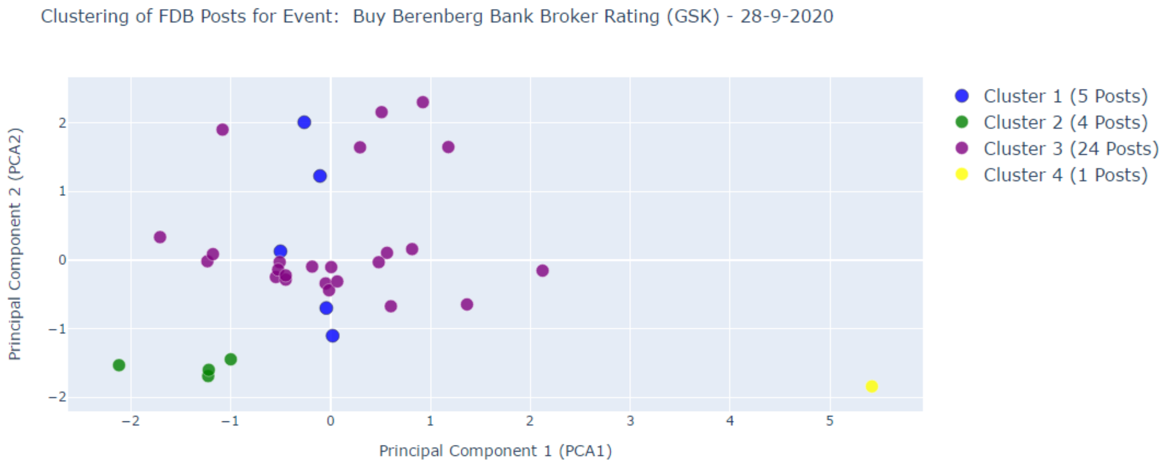
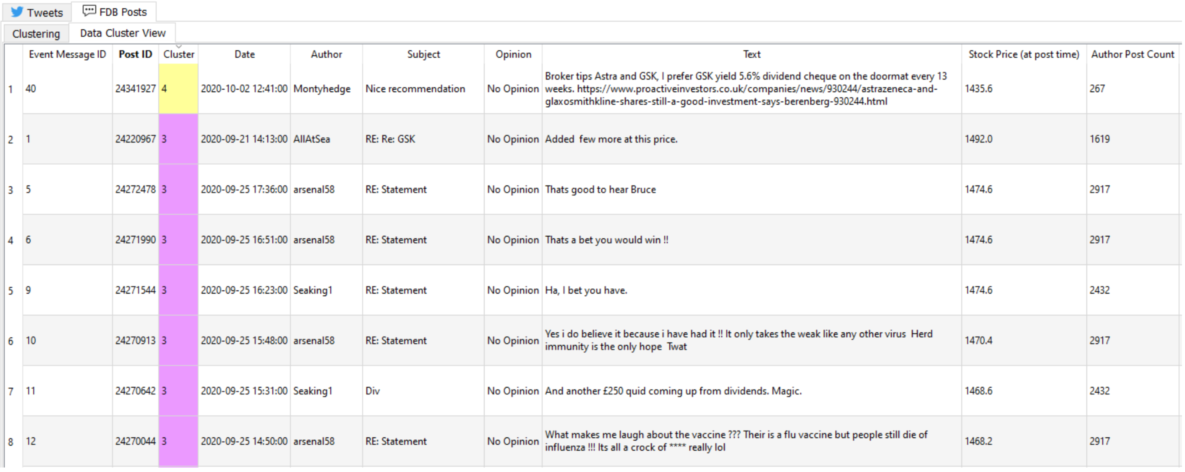


FIGURE 9.14: Scenario 6 - Clustering of FDB Posts for Berenberg Buy Broker Rating - GlaxoSmithKline PLC (LON:GSK) Event 28/09/2020

Participants A, C & E cited they would be initially interested in the single outlier data point belonging to cluster 4. Participant B did not know which cluster they would refer to as irregular, and that doing any kind of analysis would take too much of their time. Participant D stated they would be interested in clusters 1 & 3 as an investor, and the outlier clusters (2 & 4) as a regulator. Participant E noted there appears to be a slight upwards correlation in the first three clusters. They also added that the yellow cluster was difficult to see initially as it blends into the background.

Participants were then asked: *Based on viewing the data point(s) for this cluster, is it clear why the ecosystem has decided to cluster the data point(s) in this way?*. Participants were shown the individual data points, along with the features used to perform the clustering (Figure 9.15).



	Event Message ID	Post ID	Cluster	Date	Author	Subject	Opinion	Text	Stock Price (at post time)	Author Post Count
1	40	24341927	4	2020-10-02 12:41:00	Montyhedge	Nice recommendation	No Opinion	Broker tips Astra and GSK, I prefer GSK yield 5.6% dividend cheque on the doormat every 13 weeks. https://www.proactiveinvestors.co.uk/companies/news/930244/astrazeneca-and-glaxosmithkline-shares-still-a-good-investment-says-berenberg-930244.html	1435.6	267
2	1	24220967	3	2020-09-21 14:13:00	AllAtSea	RE: Re: GSK	No Opinion	Added few more at this price.	1492.0	1619
3	5	24272478	3	2020-09-25 17:36:00	arsenal58	RE: Statement	No Opinion	Thats good to hear Bruce	1474.6	2917
4	6	24271990	3	2020-09-25 16:51:00	arsenal58	RE: Statement	No Opinion	Thats a bet you would win !!	1474.6	2917
5	9	24271544	3	2020-09-25 16:23:00	Seaking1	RE: Statement	No Opinion	Ha, I bet you have.	1474.6	2432
6	10	24270913	3	2020-09-25 15:48:00	arsenal58	RE: Statement	No Opinion	Yes i do believe it because i have had it !! It only takes the weak like any other virus Herd immunity is the only hope Twat	1470.4	2917
7	11	24270642	3	2020-09-25 15:31:00	Seaking1	Div	No Opinion	And another £250 quid coming up from dividends. Magic.	1468.6	2432
8	12	24270044	3	2020-09-25 14:50:00	arsenal58	RE: Statement	No Opinion	What makes me laugh about the vaccine ??? Their is a flu vaccine but people still die of influenza !!! Its all a crock of **** really lol	1468.2	2917

FIGURE 9.15: Scenario 6 - Data points View

All participants with the exception of E observed that the sole data point in the yellow cluster was distinguishable by the fact it was the only post to contain a URL - indicating the post has supplied supplementary information. Participant C noted the presence of a URL would warrant further investigation (e.g. is it to a reputable source and relates to the FDB post topic?). Participant D added that this post also contains three recommendations (upvotes by other FDB users), differing from data points in other clusters.

In respect to the green cluster, participant B noted that the number of posts made by the forum users in this cluster seems to be the distinguishable feature. Participant D noted data points within cluster 3 also appeared to have a high post count. Participant E offered additional insight into the yellow outlier; the post was made whilst the stock price was at its lowest compared to the other posts.

Scenario Summary

The clustering of FDB posts for an event can provide different insights to that of the tweet clustering. Firstly, FDB posts are not constrained with the same character limitations that tweets must abide by. This naturally leads to FDB posts containing less abbreviated words and slang which is dominant on Twitter. Secondly, FDBs are governed by rules that posters agree to abide by when posting on the forum, which means posts will often be less ”spammy” and more relevant to the company’s operations.

9.5.7 Closing Remarks

Once the final scenario had been discussed, participants were then asked two closing questions aimed to capture any other feedback or comments they wished to share. The first question was: *Do you believe the tools within this ecosystem contribute to the effective monitoring of stock market discussion?*

Four of the five participants (all except Participant B) agreed that the tools within the ecosystem do contribute to the effective monitoring of stock market discussion. Participant B stated that the most useful tool was the filtering aspect provided by the cashtag collision classifier, but was not convinced it was needed to monitor tweets. Participants C, D & E stated that the most useful tool was the filtering aspect of resolving cashtag collisions.

Finally, participants were asked: *Do you have any other questions or comments about the ecosystem not covered in the scenarios?*

All of the participants (with the exception of Participant B) agreed that the most significant contribution was undoubtedly the filtering aspect of the ecosystem. Participant B added ”If I was Twitter, I would adopt accepted methodologies of labelling tweets, such as techniques used by Bloomberg and Financial Times – although these companies may have copyrighted their way of expressing the company/exchange”. Participant A expressed that the ecosystem makes a lot of sense from a regulatory standpoint for identifying misbehaviour - as everyone participating in financial markets should have access to the same information, and no one should have an advantage over another.

Closing Remarks Summary

All of the tools evaluated in these interviews can operate independently of each other. However, they also complement each other - attempting to cluster all tweets for a company without resolving the colliding cashtags will undoubtedly leave to incorrect analyses. The credibility aspect of this work has proved to be the most contentious in terms of feedback obtained from the participants.

9.6 Chapter Summary

This chapter has detailed the evaluation of the ecosystem through the use of interviews with expert participants. The major contributions of the ecosystem - the filtering of tweets containing cashtag collisions, the credibility assessment of tweets, and the clustering capabilities - were all evaluated. In respect to the research question that was to be addressed this chapter - "Can a smart data ecosystem, through visualisation tools, assist a user in establishing the significance of detected irregularities?": Four of the five participants agree that the ecosystem contributes to the effective monitoring of financial market monitoring in respect to the monitoring of discussion. The answers and discussion generated from the evaluative interviews support the hypothesis raised at the start of this chapter - that an ecosystem that offers visualisation tools to support the analysis of irregularities could be beneficial from a regulatory standpoint.

Chapter 10

Conclusions and Future Work

10.1 Overview

The research presented in this thesis concerned the design and development of a novel multi-layered, Smart Data Ecosystem (SDE) for the monitoring of stock discussion relating to the London Stock Exchange. Along the way, several challenges associated with monitoring the discussion taking place relating to stocks - and the context of such discussion - was explored and resolved. The principal challenge of this research related to the collection of stock tweets, as such tweets can contain a cashtag that can refer to multiple companies listed on different stock exchanges. This chapter begins by reviewing the key contributions of the work (Section 10.2), and then reviews how the research questions were addressed (Section 10.3), with Section 10.4 proposing avenues for future work.

10.2 Review of Contributions

This section will revisit the contributions outlined in Section 1.6.

10.2.1 Smart Data Ecosystem

The principal contribution of this research is the SDE, which monitors multiple communication channels to attempt to identify potentially irregular behaviour on the part of investors. The SDE houses various tools (each of which will be discussed shortly), that can operate independently, or can be combined to produce more accurate results - e.g. the filtering of non-LSE tweets is undertaken through the use of the cashtag collision classifier (Section 6.7), which can then be clustered using the approaches outlined in Chapter 8. Each of these tools will now be discussed, including how they contribute to the SDE.

10.2.2 Data Fusion Model

The data fusion model presented in Section 5.7 provides the basis of the data layer, in which data feeders for different data sources and communicative platforms feed data into the data fusion model. As the data makes its way through the data fusion model, issues such as establishing if the tweet refers to the stock exchange or not are resolved (Level 3). Once cashtag collisions have been resolved, data pertaining to specific companies is stored in time-slice windows of a single day, in addition to being stored in event-specific documents. A day is a typical time unit for dividing stock market data - each stock has an opening and close price for a trading day, and the performance of a stock index is often reported by reporting the opening and closing price at the start/end of the trading day. This data fusion model contributes to the SDE by providing synchronicity for the different data sources, and alleviates issues such as differences in API timestamps for the different data sources, accounting for time zone differences between such APIs, and resolving cashtag collisions.

10.2.3 Resolving Cashtag Collisions

A novel methodology for the resolution of cashtag collisions on Twitter (Chapter 6) is one of the key contributions of this work. This issue had yet to be addressed within the literature until our paper in Evans, Owda, Crockett, and Ana Fernandez Vilas (2019). The methodology for resolving cashtag collisions (Section 6.6) proposed two sets of features: (1) a sparse vector of the tweet text, and (2) a sparse vector of the tweet text combined with frequency keyword counts of terms within the tweet that are also used on the LSE-variant Reuters page, in addition to keywords being used by investors on the London South East forum. This element of the research found that the inclusion of features derived from company-corpora assisted in the detection and resolution of cashtag collisions.

10.2.4 Assessing the Credibility of Financial Stock Tweets

Chapter 7 presented a methodology - and an experiment to validate the methodology - for assessing the credibility of financial stock tweets. Two sets of features were proposed: (1) general features that can be found, or engineered, in any tweet, and (2) financial features that can be found, or engineered, in tweets containing at least one cashtag of a stock-listed company. In total, 44 general features and 15 financial features were considered in the training of the credibility classifiers, with the best performing classifier (with respect to the ROC-AUC metric) being a Random Forest classifier trained on 25 general features and 12 financial features.

10.2.5 Detection of potential irregularities

Chapter 8 presented the detection capability of the SDE, which take the form of clustering algorithms that look at significant events in a company's operations. The clustering can be undertaken on all events for a company (Section 8.5), the tweets for a company event (Section 8.6), or the FDB posts for an event (Section 8.7).

10.3 Review of Research Questions

Section 1.5 outlined four research questions to be addressed in this research. These four research questions, and how the thesis has addressed each, are summarised below.

1. **Can a smart data ecosystem, utilising machine learning classifiers, classify social media posts with respect to their credibility? (Chapter 7)**

   This research question was addressed in Chapter 7. Firstly, tweets relating to companies in Appendix E were collected and filtered to remove non-LSE tweets, using the methodology outlined in Section 6.6. A subset of tweets was then selected to annotate for credibility, in which a three-label system was ultimately adopted: (0) not credible, (1) ambiguous, and (2) credible. Multiple annotators (the main annotator and three others) annotated the credibility of a subsample of the tweets, and their annotations were compared to see if the annotators shared a high level of agreement as to what constitutes tweet credibility. In regards to training classifiers, two feature sets were proposed and used in the training of classifiers - general features that can be found (or engineered from) any tweet - regardless of subject matter - and financial features that can be found (or engineering from) financial stock tweets (tweets that contain at least one cashtag). Before the training of classifiers commenced, feature selection techniques were adopted to identify features that offer no, or very little, informative power to the classifiers. Two sets of classifiers were then trained: the first set of classifiers were trained solely on the general features, with the second set trained on both general and financial features. The best performing classifier, in respect to the ROC-AUC metric, was the Random Forest classifier (AUC: 94.3) trained on both sets of features. However, the Random Forest classifier required 37 features in total (25 general features, and 12 financial features), whereas the k-nearest neighbours classifier trained on both feature sets required only 9 features (7 general features, and 2 financial features) and yielded an AUC score of 93.6.

2. **Can a smart data ecosystem be used to monitor a variety of communication channels for irregular behaviour? (Chapter 8)**

   This research question was addressed in Chapters 8 & 9. Firstly, the detection capabilities of the SDE were outlined and discussed in Chapter 8, in which three clustering approaches were introduced: (1) event-based clustering (Section 8.5, (2) tweet-based clustering within an event (Section 8.6), and (3) FDB-based clustering within an event (Section 8.7). Chapter 9 then delved deeper into the detection layer capability by reporting on the results and discussion of conducting qualitative interviews with five financial market experts, centring on specific tools within the SDE via scenario-based questions. Four of the five participants interviewed agreed that the SDE does contribute to the effective monitoring of financial market discussion.

3. **Can a smart data ecosystem, utilising clustering algorithms, identify irregular days and events with respect to posting activity? (Chapter 8)**

This research question was addressed in Chapter 8, whereby the detection capabilities of the SDE were introduced, utilising the popular k-means clustering algorithm. Three types of clustering approaches are adopted by the detection layer: (1) the clustering of two-week time windows (events) in respect to features such as the number of tweets and FDB posts in the window, (2) the clustering of tweets in an event window, and (3) the clustering of FDB posts in an event window. These detection capabilities were then evaluated through a set of qualitative interviews, which leads to the next research question.

4. **Can a smart data ecosystem, through visualisation tools, assist a user in establishing the significance of detected irregularities? (Chapter 9)**

   This research question was addressed in Chapter 9, in which the five financial markets experts were asked various questions relating to the clustering capability of the SDE. They were asked specific questions regarding the data points belonging to the clusters as a result of plotting the two principal components of the clustering outputs. Ultimately, four of the five experts agreed that the SDE did contribute to the effective monitoring of financial market communication.

   On the outset of this research, the issues of cashtag collisions was not immediately known, and if indeed the phenomenon of colliding cashtags was not identified or resolved, it would undoubtedly lead to the clustering of such tweets, and the visualisation of the clusters, to be susceptible to incorrect analyses and interpretation. Chapter 6 therefore contributes to this research question by aiding the visualisation process of the clustering output by ensuring non-LSE and cryptocurrency tweets are not included in the clustering and visualisation process.

10.4 Future Work

The research presented in this thesis has explored the challenges associated with collecting financial market data at a large scale, the fusion of such disparate data source, and using time-slice and event time-windows for detecting irregularities.

This section will now explore avenues of potential future research relating to this research project.

10.4.1 Large-scale Tweet Collection

One of the principal shortcomings of the data collection aspect of this research is that the collection of tweets is limited to 1% tweets, and enterprise-level APIs that allow larger-scale data collection are expensive. Recently, Twitter has announced a specialised API for academic research[1], that allows enterprise-level collection of tweets (up to 10 million tweets a month), far surpassing the free version which is limited in the number of search filters that can be applied.

10.4.2 Automatic Detection of Irregularities

One of the principal limitations of the presentation and decision layer is that the clustering is performed manually on a company and event-specific basis. This means that the user must pre-select which events to cluster the FDB posts and tweets for, meaning the process is independent of the SDE. The ability to perform this clustering in the background and then have some way of reporting which clustering outputs would be of particular interest to a regulator would make a substantial contribution to the effective monitoring of financial market discussion.

10.4.3 Additional Events

The two-week events generated in Section 8.3 focus solely on broker analyst ratings that fall within a buy or sell category. Naturally, by generating more events to hone in on other periods of discussion activity could provide more insights into investor behaviour (e.g. Do appointments of new Chief Executive Officers lead to more irregular activity being detected when compared to buy and sell analyst ratings?).

Examples of other events could include:

- **Regulatory News Service releases** - RNS announcements are made by a company to in order to inform the investors and other market participants about

[1] https://developer.twitter.com/en/products/twitter-api/academic-research

the company's operations. This could include addressing speculation or rumours that are circulating in the media, or the announcement of a new CEO/CFO or other significant board appointment.

- **Pre-defined price movements** - Each company could be assigned a company-specific price threshold which causes an event to be generated if such a threshold is met. The activity before and after this event could provide insight such as if certain investors had insider knowledge before the price movement happened.
- **Posting/tweet activity** - the volume of FDB posts and/or tweets could itself serve as a mechanism for generating events, whereby every company has thresholds that are constantly being monitored and adjusted by the ecosystem.
- **External events** - events that occur outside the financial markets - such as natural disasters and disease outbreaks that lead to financial markers destabilising.

Appendix A

Data Fusion Paper

Intelligent Systems Conference 2018
6-7 September 2018 | London, UK

Big Data Fusion Model for Heterogeneous Financial Market Data (FinDF)

Lewis Evans, Majdi Owda, Keeley Crockett
School of Computing, Mathematics & Digital Technology
Manchester Metropolitan University M41 5GD UK
Manchester, UK
{l.evans, m.owda, k.crockett}@mmu.ac.uk

Ana Fernández Vilas
I&C Lab. AtlantTIC Research Centre
University of Vigo. 36310
Pontevedra, Spain
avilas@det.uvigo.es

***Abstract*— The dawn of big data has seen the volume, variety, and velocity of data sources increase dramatically. Enormous amounts of structured, semi-structured and unstructured heterogeneous data can be garnered at a rapid rate, making analysis of such big data a herculean task. This has never been truer for data relating to financial stock markets, the biggest challenge being the 7Vs of big data which relate to the collection, pre-processing, storage and real-time processing of such huge quantities of disparate data sources. Data fusion techniques have been adopted in a wide number of fields to cope with such vast amounts of heterogeneous data from multiple sources and fuse them together in order to produce a more comprehensive view of the data and its underlying relationships. Research into the fusing of heterogeneous financial data is scant within the literature, with existing work only taking into consideration the fusing of text-based financial documents. The lack of integration between financial stock market data, social media comments, financial discussion board posts and broker agencies means that the benefits of data fusion are not being realised to their full potential. This paper proposes a novel data fusion model, inspired by the data fusion model introduced by the Joint Directors of Laboratories, for the fusing of disparate data sources relating to financial stocks. Data with a diverse set of features from different data sources will supplement each other in order to obtain a Smart Data Layer, which will assist in scenarios such as irregularity detection and prediction of stock prices.**

Keywords*— Big Data, Data Fusion, Heterogeneous Financial Data

I. Introduction

The ineluctable growth of heterogeneous financial data sources relating to financial stocks poses a serious challenge to researchers and regulators who attempt to analyse stock market discussions and prices for a variety purposes such as detecting possible irregular behaviour [1][2]. With the advent of social media, financial discussion boards (FDBs), and traditional news media dissemination, investors have an almost endless amount of communication channels to make use of for executing well-informed investments [3]. The analysis of such communication is difficult to undertake, due to the many problems associated with big data within the financial market domain [1][4]. Big data is defined as "data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data" [4].

There exists a myriad of studies on the Vs of big data, the first instance being the consideration of volume, velocity, and variety [5], since then there have been extensions to the Vs of big data, including the 4Vs[6], 5Vs[7], 7Vs[8], and more recently, a 42V approach to big data has been proposed [9]. For our study on financial stock markets, we adopt the 7Vs conceptual model of big data (volume, variety, velocity, variability, veracity, value and visualisation), as these seven are clearly distinguishable in the field of financial stock markets [4]. The increasing number of Vs in source data, the more complex the fusion process will be in order to produce Smart Data.

Data fusion has been a well-established practice for managing heterogeneous data sources through the use of associating and combining data sources together [10][11]. Several models proposed for the fusion of data include the model proposed by the Joint Directors of Laboratories (JDL) [12] and the Dasarathy model [13]. These models, however, have been outdated due to their emphasis on specific domains and applications, often needing to be revised and adapted based on the specific fusion task [14].

Limited research has been undertaken on the fusion of financial data sources, in this paper we coin the term FinDF to refer to the fusing of financial data sources. Existing fusion techniques do not consider more than two data sources, and focus on Securities and Exchange Commission (SEC) filings (which are only available for stocks listed on US exchanges such as the NYSE or NASDAQ) along with other text-based document filings [15]. The existing challenges of FinDF lie in the fact that each of these financial data sources have a different origin, their contents will often be distributed over a variety of websites and vary dramatically in terms of their structure and intent. As existing research focuses primarily on integrating textual documents, there is an opportunity to improve upon existing methodologies by establishing data fusion techniques which take into account data sources such as social media comments, financial discussion board posts, broker agency ratings and stock market data.

This paper proposes a novel data fusion model to address the fusion of financial data from multiple source environments, providing a solution for the current challenges of data association from multiple environments, namely *how* to fuse such data. The proposed model will approach the fusion task from two dimensions; (1) fusing the different data sources together based on time-slice windows and (2) the company in which the data corresponds to.

This paper is organised as follows: Section II looks at the related work on data fusion, including its use in various fields and how the JDL model has inspired existing fusion tasks. Section III introduces some of the financial data sources which are used by investors to discuss stocks and make investment decisions. Section IV explores the challenges of big data in relation to financial markets, and how the 7 Vs of big data are dominant within the field of financial markets. Section V presents the proposed FinDF model for the fusing of financial data sources. Section VI explores the future work which could be performed as a result of this research, in addition to drawing a conclusion in relation to how the FinDF model addresses some of the challenges of big data within the financial market domain.

II. Related Work

A. Data Fusion

Several definitions exist within the literature for the term data fusion. The first definition being coined by Hall and Llinas [16]: "data fusion techniques combine data from multiple sensors, and related information from associated databases, to achieve improved accuracies and more specific inferences that could be achieved by the use of a single sensor alone".

The terms data fusion and information fusion are often used synonymously; there is, however, a distinction which should be made. The term data fusion is used to refer to fusing raw data (data which is obtained directly from a source with no pre-processing or cleaning being carried out), whereas the term information fusion is used to refer to the fusion of data which is already processed in some way [17]. Regardless of the term used, data and information fusion techniques are used to enhance knowledge discovery [18].

There exist a considerable number of challenges associated with the fusion of data sources, many of these challenges stem from the disparity of how different data is structured [19]. The most notable challenges, outlined by [20], include:

1) Disparate Data

The input data which is provided to a data fusion model will most often be generated by a variety of sources such as humans (e.g. textual comments), APIs (e.g. time-stamped sequential data), scraping (e.g. textual content). Fusion of such heterogeneous data in order to construct a comprehensible and accurate view of the overall picture is a challenging task in itself.

2) Outliers and False data

Noise and impreciseness of data can be found in almost all sources of data. A data fusion algorithm should be able to take measures against outliers which are presented to it and take appropriate action accordingly as part of the fusion process.

3) Data Conflict

Data fusion algorithms must be able to treat conflicting data with great care, being careful not to simply discard it, but to provide a means of cross-checking the data across the different sources.

4) Imperfection of Data

Data will often be affected by some element of impreciseness, a data fusion model should be able to express such imperfections and make a decision such as whether or not to discard such data, or fuse the data and accept the risk of imperfect data fusion.

5) Out of Sequence Data

Data which is inputted into a data fusion model will often be organised in discrete pieces which feature a corresponding timestamp, detailing its time of origin. Undoubtedly, the different input sources may be out of sequence due to varying time-zones in which the data is collected from, including factors such as daylight-saving time.

6) Data Association

Associating multiple entities into groups is the most significant problem of the data fusion process. It can be seen as trying to establish hidden or secret relationships between entities which may not appear to be immediately apparent.

7) Data Collection

As is the case with many web 2.0 technologies, APIs are often provided for the unified collection of data. However, not all sources provide such a convenient way of collecting data, meaning techniques such as web scraping will need to be utilised for data collection.

B. Fields Utilising Data Fusion

Data fusion has been employed successfully in a wide range of domains in order to combine multiple data sources into a unified data output [21]. Table I lists several fields in which data fusion has been adopted to improve the accuracy of analysing multiple data sources.

The success of data fusion in these domains through the use of fusing different data relating to the same objects for better observations make it an attractive option for combining financial stock market data.

Although work has been undertaken which integrates market data with financial news and work which considers the fusion of documents, this work does not consider the fusion of such a wide variety of disparate data sources such as social media comments, discussion board posts and broker agency ratings [22][23]. To our knowledge, there has been no work undertaken which considers the fusion of multiple disparate data sources relating to financial stock markets.

C. Data Fusion Models

There have been a number of reviews of existing data fusion models and architectures in recent years [17][20]. Existing models include the Intelligence cycle model, Boyd control loop model, Dasarathy model, and the Thompoulos model [24]. Although there have been several proposals of data fusion models over the years, none have become more widely adopted as the JDL model [25], which will now be overviewed in detail.

1) JDL Model

Initially proposed by the U.S Joint Directors of Laboratories (JDL) and the U.S Department of Defense (DoD) in 1985 [24, p. 111], the JDL model is considered the seminal model for data fusion tasks [26]. The JDL model (Fig 1) is comprised of five processing levels, a database management system (DBMS), human interaction, and a data bus which connects all of these components together [27].

a) Level 0 – Source Pre-processing

The lowest layer present in the JDL model involves reducing the volume of the data using data cleaning techniques, addressing missing values, and maintaining useful information for the higher-level processes.

b) Level 1 – Object Refinement

At this low-level of the fusion model, data is aligned to objects in order to allow statistical estimation, and to permit common data processing [26][28].

TABLE I FIELDS UTILISING DATA FUSION

Field	Description	Refs
Forensics - Network Intrusion Detection Systems (IDS)	Complementing evidence and artifacts from different layers of a computer or devices to create a complete picture of what events occurred during a reactive forensic investigation. The proposed model (based on the JDL model) can successfully reduce false positive alarms generated by IDS and improve the detection of unknown threats.	[67]
Military – Unmanned Aerial Vehicles (UAV)	Detection of threats based on multi-sensor multi-source data fusion. The proposed model (also based on the JDL) aimed to enhance the situation awareness of the UAV (human) operators by providing a model supporting the detection of threats based on different data sources fused together.	[68]
Navigation Systems	Beacons used for navigation systems and emergencies are highly susceptible to noise, frequency shifts and measurement errors. The adoption of data fusion was able to reduce packet error rate from beacons and sensors from 70% to 4.5%.	[64]
Track monitoring from multiple in-service trains	Monitoring of a rail-track network to ensure safety of its users and to reduce maintenance costs by early detection of faults. The proposed model, which fused position data from trains, and track data (vibrations), indicated that fusing data helped in the detection of track changes, resulting in early detection of track faults.	[69]
Geosciences – Habitat Mapping	Data combined from multiple sources (hyperspectral, aerial photography, and bathymetry data) was utilised for the purposes of mapping and monitoring of the benthic habitat in the Florida Keys.	[70]

c) Level 2 – Situation Refinement

This level deals with the relationships between objects and observed events, attempting to provide a contextual description between the relationships [27][29].

d) Level 3 – Threat Refinement

The fusion process of this level attempts to create data for future predictions. The output of which is prediction data which can be stored for further analysis or acted upon [21].

e) Level 4 – Process Refinement

The monitoring of system performance, including handling real time constraints is addressed at this level [29]. This level of the data fusion model does not perform any data processing operations, as it is more focused on identifying information required for data fusion improvement [30][31].

f) Support Database

The support database of the JDL model serves as a data repository in which raw data is stored to facilitate the fusion process. [32]

g) Fusion Database

At the conclusion of the data fusion process, fused data is stored within the fusion database, to be used for future analysis tasks.

2) JDL Model Revisions

The original JDL data fusion model was incepted to provide a process flow for sensor and data fusion [14]. As a result of the JDL model being over thirty years old, it has been revised over the years to address specific data fusion challenges. Despite the popularity of the JDL model, it has been subject to scrutiny due to being tuned primarily for military applications and being too restrictive [20]. Revisions to the JDL model in 1999 by [33], involved a redefined model which attempted to steer away from a model which, at the time, was tailored primarily for military applications, which was the case for many data fusion tasks at that period [34].This revision to the JDL model revolved primarily around redefining the Threat Refinement process; as the concept of "threats" does not exist to such an extent as it does in the military domain. Steinberg, Bowman, and White [33] redefined the *Threat Refinement* level as *Impact Assessment*, as impact is considered an umbrella-term which, unlike threat refinement, is not restricted to specific domains.

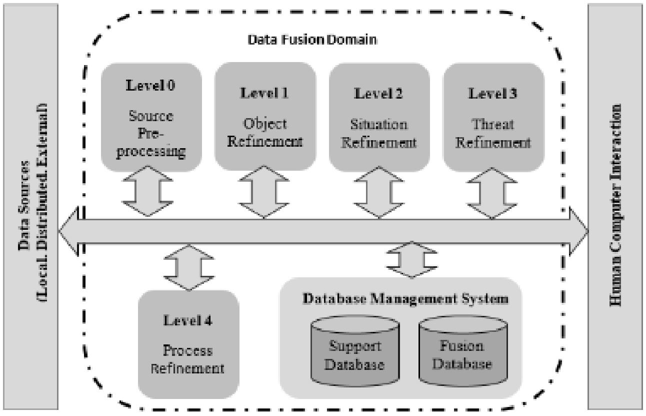


Fig 1 JDL Data Fusion Model

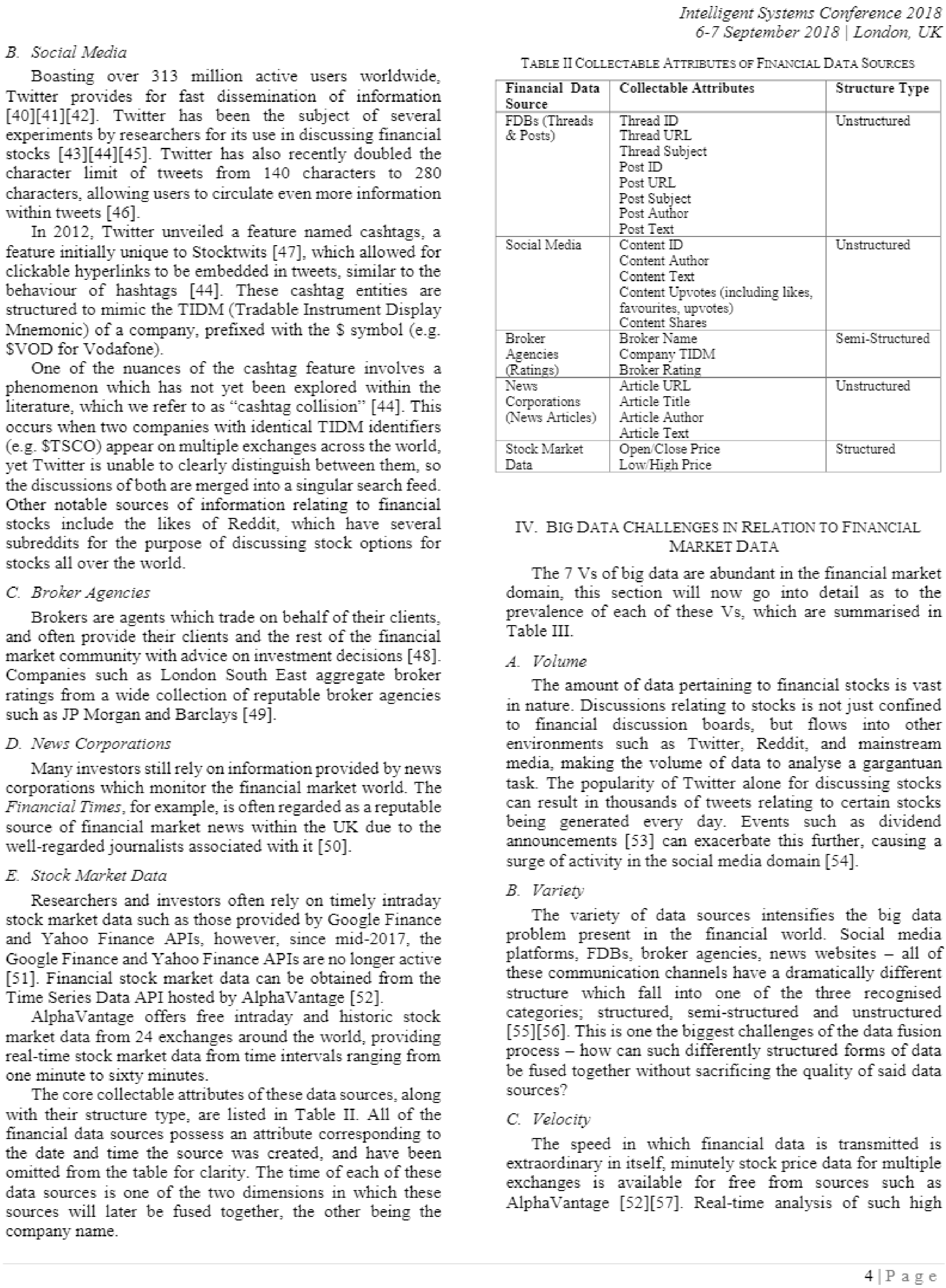
Further revisions and extensions to the JDL model were proposed in 2004 by [35]. Proposals in this paper involved extending the model to include the previous remarks on issues relating to quality control, reliability, and the consistency in data fusion processing.

III. FINANCIAL DATA SOURCES

Investors have a plethora of information sources when it comes to researching and discussing stock options. The data fusion model we propose will utilise sources from a variety of environments. In this section, we will detail the data sources which will be fused by the data fusion model.

A. Financial Discussion Boards (FDBs)

During the early 2000s, the emergence of financial discussion boards such as Yahoo! Finance and Raging Bull provided two of the most prominent messaging boards on the internet [36]. FDBs provide an unprecedented opportunity for investors to invest, debate, and exchange information on stocks, often expressing their own individual opinion, and often having no prior social connections to other users [37]. FDBs are often specific to certain stock markets, Interactive Investor and London South East, for example, provide a platform for investors to discuss stocks which float on the London Stock Exchange, offering a separate discussion board for each stock [38]. Existing work undertaken by [39] has utilised this data source for the purpose of highlighting potential irregularities through the use of information extraction (IE).



B. Social Media

Boasting over 313 million active users worldwide, Twitter provides for fast dissemination of information [40][41][42]. Twitter has been the subject of several experiments by researchers for its use in discussing financial stocks [43][44][45]. Twitter has also recently doubled the character limit of tweets from 140 characters to 280 characters, allowing users to circulate even more information within tweets [46].

In 2012, Twitter unveiled a feature named cashtags, a feature initially unique to Stocktwits [47], which allowed for clickable hyperlinks to be embedded in tweets, similar to the behaviour of hashtags [44]. These cashtag entities are structured to mimic the TIDM (Tradable Instrument Display Mnemonic) of a company, prefixed with the $ symbol (e.g. $VOD for Vodafone).

One of the nuances of the cashtag feature involves a phenomenon which has not yet been explored within the literature, which we refer to as "cashtag collision" [44]. This occurs when two companies with identical TIDM identifiers (e.g. $TSCO) appear on multiple exchanges across the world, yet Twitter is unable to clearly distinguish between them, so the discussions of both are merged into a singular search feed. Other notable sources of information relating to financial stocks include the likes of Reddit, which have several subreddits for the purpose of discussing stock options for stocks all over the world.

C. Broker Agencies

Brokers are agents which trade on behalf of their clients, and often provide their clients and the rest of the financial market community with advice on investment decisions [48]. Companies such as London South East aggregate broker ratings from a wide collection of reputable broker agencies such as JP Morgan and Barclays [49].

D. News Corporations

Many investors still rely on information provided by news corporations which monitor the financial market world. The *Financial Times*, for example, is often regarded as a reputable source of financial market news within the UK due to the well-regarded journalists associated with it [50].

E. Stock Market Data

Researchers and investors often rely on timely intraday stock market data such as those provided by Google Finance and Yahoo Finance APIs, however, since mid-2017, the Google Finance and Yahoo Finance APIs are no longer active [51]. Financial stock market data can be obtained from the Time Series Data API hosted by AlphaVantage [52].

AlphaVantage offers free intraday and historic stock market data from 24 exchanges around the world, providing real-time stock market data from time intervals ranging from one minute to sixty minutes.

The core collectable attributes of these data sources, along with their structure type, are listed in Table II. All of the financial data sources possess an attribute corresponding to the date and time the source was created, and have been omitted from the table for clarity. The time of each of these data sources is one of the two dimensions in which these sources will later be fused together, the other being the company name.

TABLE II COLLECTABLE ATTRIBUTES OF FINANCIAL DATA SOURCES

Financial Data Source	Collectable Attributes	Structure Type
FDBs (Threads & Posts)	Thread ID Thread URL Thread Subject Post ID Post URL Post Subject Post Author Post Text	Unstructured
Social Media	Content ID Content Author Content Text Content Upvotes (including likes, favourites, upvotes) Content Shares	Unstructured
Broker Agencies (Ratings)	Broker Name Company TIDM Broker Rating	Semi-Structured
News Corporations (News Articles)	Article URL Article Title Article Author Article Text	Unstructured
Stock Market Data	Open/Close Price Low/High Price	Structured

IV. BIG DATA CHALLENGES IN RELATION TO FINANCIAL MARKET DATA

The 7 Vs of big data are abundant in the financial market domain, this section will now go into detail as to the prevalence of each of these Vs, which are summarised in Table III.

A. Volume

The amount of data pertaining to financial stocks is vast in nature. Discussions relating to stocks is not just confined to financial discussion boards, but flows into other environments such as Twitter, Reddit, and mainstream media, making the volume of data to analyse a gargantuan task. The popularity of Twitter alone for discussing stocks can result in thousands of tweets relating to certain stocks being generated every day. Events such as dividend announcements [53] can exacerbate this further, causing a surge of activity in the social media domain [54].

B. Variety

The variety of data sources intensifies the big data problem present in the financial world. Social media platforms, FDBs, broker agencies, news websites – all of these communication channels have a dramatically different structure which fall into one of the three recognised categories; structured, semi-structured and unstructured [55][56]. This is one the biggest challenges of the data fusion process – how can such differently structured forms of data be fused together without sacrificing the quality of said data sources?

C. Velocity

The speed in which financial data is transmitted is extraordinary in itself, minutely stock price data for multiple exchanges is available for free from sources such as AlphaVantage [52][57]. Real-time analysis of such high

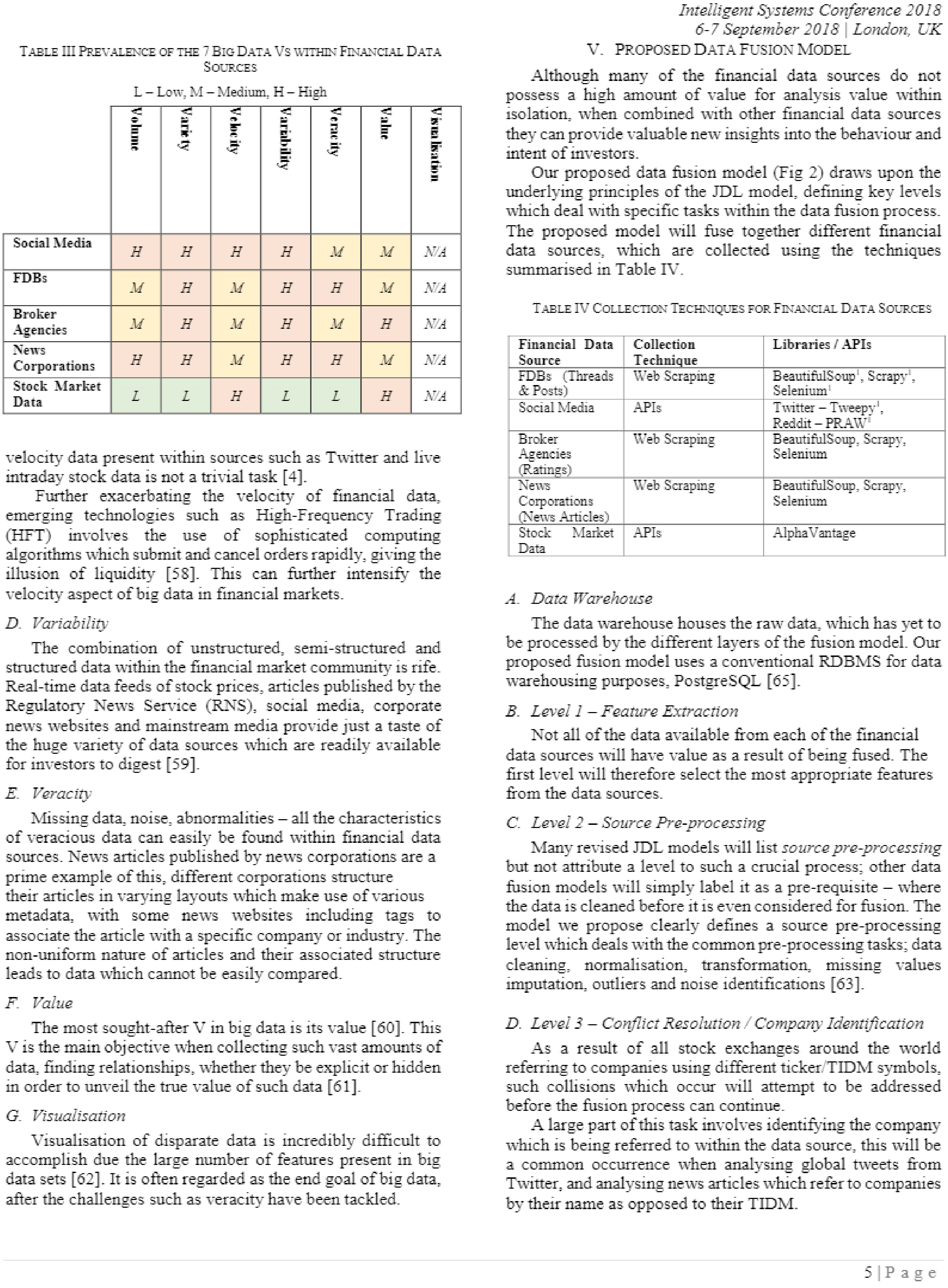


TABLE III PREVALENCE OF THE 7 BIG DATA VS WITHIN FINANCIAL DATA SOURCES

L – Low, M – Medium, H – High

	Volume	Variety	Velocity	Variability	Veracity	Value	Visualisation
Social Media	*H*	*H*	*H*	*H*	*M*	*M*	*N/A*
FDBs	*M*	*H*	*M*	*H*	*H*	*M*	*N/A*
Broker Agencies	*M*	*H*	*M*	*H*	*M*	*H*	*N/A*
News Corporations	*H*	*H*	*M*	*H*	*H*	*M*	*N/A*
Stock Market Data	*L*	*L*	*H*	*L*	*L*	*H*	*N/A*

velocity data present within sources such as Twitter and live intraday stock data is not a trivial task [4].

Further exacerbating the velocity of financial data, emerging technologies such as High-Frequency Trading (HFT) involves the use of sophisticated computing algorithms which submit and cancel orders rapidly, giving the illusion of liquidity [58]. This can further intensify the velocity aspect of big data in financial markets.

D. *Variability*

The combination of unstructured, semi-structured and structured data within the financial market community is rife. Real-time data feeds of stock prices, articles published by the Regulatory News Service (RNS), social media, corporate news websites and mainstream media provide just a taste of the huge variety of data sources which are readily available for investors to digest [59].

E. *Veracity*

Missing data, noise, abnormalities – all the characteristics of veracious data can easily be found within financial data sources. News articles published by news corporations are a prime example of this, different corporations structure their articles in varying layouts which make use of various metadata, with some news websites including tags to associate the article with a specific company or industry. The non-uniform nature of articles and their associated structure leads to data which cannot be easily compared.

F. *Value*

The most sought-after V in big data is its value [60]. This V is the main objective when collecting such vast amounts of data, finding relationships, whether they be explicit or hidden in order to unveil the true value of such data [61].

G. *Visualisation*

Visualisation of disparate data is incredibly difficult to accomplish due the large number of features present in big data sets [62]. It is often regarded as the end goal of big data, after the challenges such as veracity have been tackled.

V. PROPOSED DATA FUSION MODEL

Although many of the financial data sources do not possess a high amount of value for analysis value within isolation, when combined with other financial data sources they can provide valuable new insights into the behaviour and intent of investors.

Our proposed data fusion model (Fig 2) draws upon the underlying principles of the JDL model, defining key levels which deal with specific tasks within the data fusion process. The proposed model will fuse together different financial data sources, which are collected using the techniques summarised in Table IV.

TABLE IV COLLECTION TECHNIQUES FOR FINANCIAL DATA SOURCES

Financial Data Source	**Collection Technique**	**Libraries / APIs**
FDBs (Threads & Posts)	Web Scraping	BeautifulSoup[1], Scrapy[1], Selenium[1]
Social Media	APIs	Twitter – Tweepy[1], Reddit – PRAW[1]
Broker Agencies (Ratings)	Web Scraping	BeautifulSoup, Scrapy, Selenium
News Corporations (News Articles)	Web Scraping	BeautifulSoup, Scrapy, Selenium
Stock Market Data	APIs	AlphaVantage

A. *Data Warehouse*

The data warehouse houses the raw data, which has yet to be processed by the different layers of the fusion model. Our proposed fusion model uses a conventional RDBMS for data warehousing purposes, PostgreSQL [65].

B. *Level 1 – Feature Extraction*

Not all of the data available from each of the financial data sources will have value as a result of being fused. The first level will therefore select the most appropriate features from the data sources.

C. *Level 2 – Source Pre-processing*

Many revised JDL models will list *source pre-processing* but not attribute a level to such a crucial process; other data fusion models will simply label it as a pre-requisite – where the data is cleaned before it is even considered for fusion. The model we propose clearly defines a source pre-processing level which deals with the common pre-processing tasks; data cleaning, normalisation, transformation, missing values imputation, outliers and noise identifications [63].

D. *Level 3 – Conflict Resolution / Company Identification*

As a result of all stock exchanges around the world referring to companies using different ticker/TIDM symbols, such collisions which occur will attempt to be addressed before the fusion process can continue.

A large part of this task involves identifying the company which is being referred to within the data source, this will be a common occurrence when analysing global tweets from Twitter, and analysing news articles which refer to companies by their name as opposed to their TIDM.

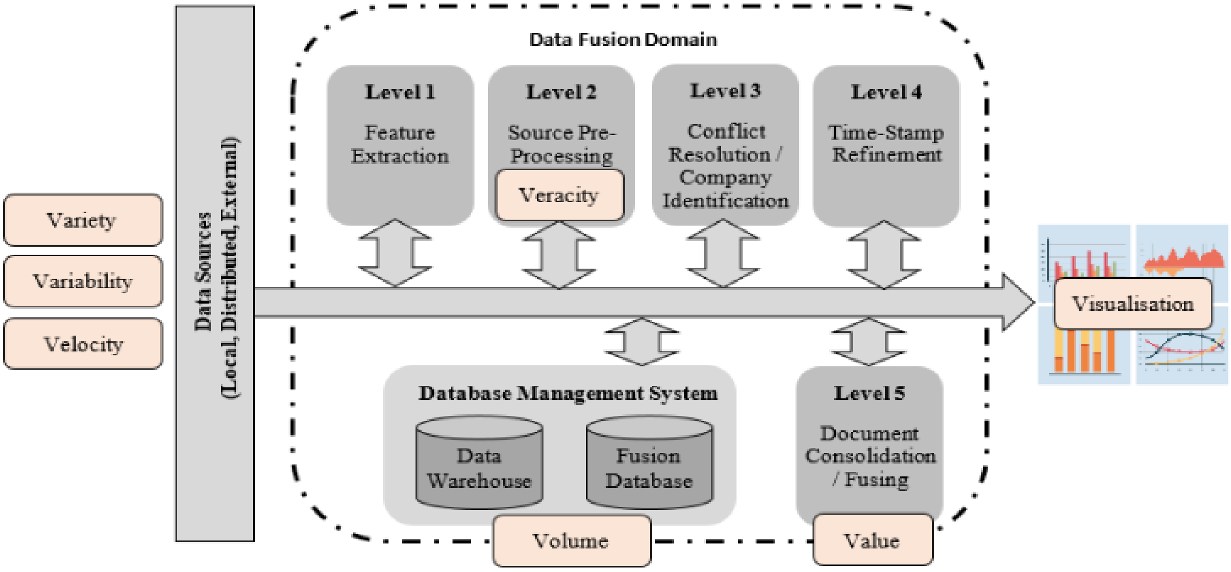


Fig 2 Proposed Financial Data Fusion (FinDF) Model

E. Level 4 – Time-Stamp Refinement

Timestamps are the determinant feature in which disparate data can be associated. Data which does not have a timestamp associated with it cannot easily be fused with other data sources [64]. This level will address inconsistent time-stamps across the different data sources, attempting to unify the data based on pre-existing time-stamps. Nuances such as daylight-saving time and time-zone differences across the different sources will also be conducted at this level.

F. Level 5 – Document Consolidation / Fusing

After the data has gone through a vigorous cleaning process and the timestamps have been aligned across the data sources, the fusion process can then continue with storing the fused data within the document-oriented fusion database. The fusing of this data is performed in accordance with pre-determined time-slice windows (for example, 15-minute intervals), and the company TIDM (ticker symbol).

G. Fusion Database

After the final fusion level has been undertaken, fused data is stored in a document-oriented fashion, allowing the fused data to be stored in a document-oriented NoSQL structure such as that supported by MongoDB [66].

VI. Conclusions and Future Work

This paper has proposed a novel data fusion model for fusing together heterogeneous data from different financial data sources. The proposed model adapted the heavily-employed JDL data fusion model for the purposes of financial data fusion.

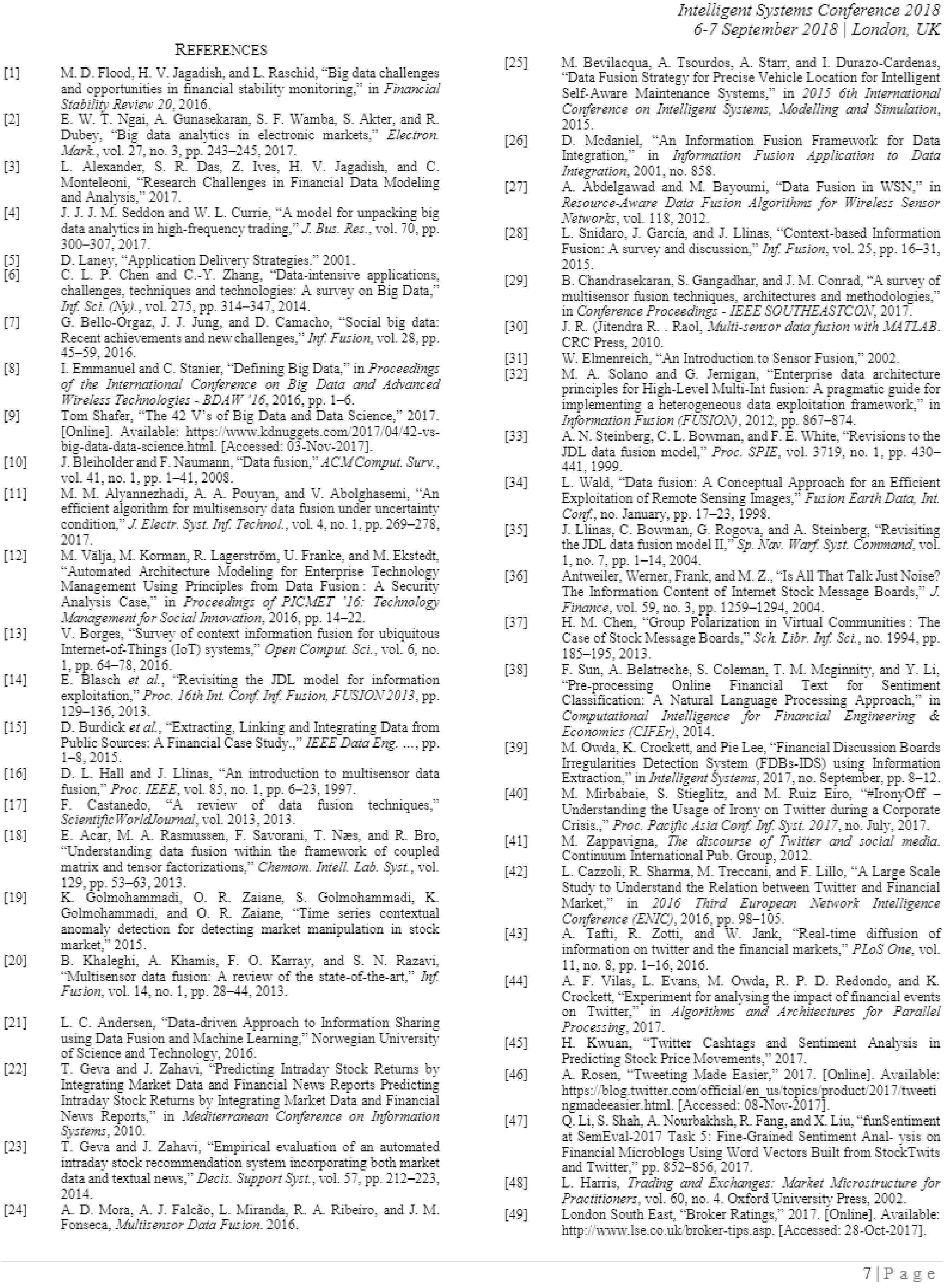
The proposed FinDF model attempts to address the challenges of working with big data within the confines of financial markets. Associating different data sources by time and company will be a challenging process when taking into consideration each of the 7 Vs of big data.

In terms of the original 3Vs (volume, variety and velocity), the fusion model will associate voluminous amounts of disparate data which is being generated at a rapid rate. Taking into consideration 2 of the other Vs (variability and veracity), these are present in the data sources in varying levels, web scraping techniques will allow us to collect data from a variety of websites, which will often be veracious in nature due to the different structure of discussion boards and other communicative websites. The last 2 Vs (value and visualisation) come after the fusion process have occurred. Although it can be argued that every data source has some inherent value in isolation, the outcome of the fusion process will allow the value to be truly apparent through the use of identifying hidden relationships between the different data sources.

Identifying the name of a company within the different data sources is also a substantial challenge which can be addressed through Natural Language Processing (NLP) techniques. The problems of cashtag collisions on Twitter could also mean that previous work undertaken could have been susceptible to incorrect analysis. HFT is also an area which requires special attention when it comes to the analysis of stock movements, such high velocity activity can make the analysis of stock market movements challenging to undertake.

The data fusion model presented in this paper will be used in the future as part of a larger multi-layered ecosystem for the monitoring of potentially irregular comments pertaining to financial stocks. This ecosystem will monitor a variety of discussion channels used by investors, in addition to news sources and utilise the data fusion model in order to amalgamate the different sources of stock information and stock prices.

To our knowledge, this is the first conceptualised model for the fusing of heterogeneous financial data sources.



REFERENCES

[1] M. D. Flood, H. V. Jagadish, and L. Raschid, "Big data challenges and opportunities in financial stability monitoring," in *Financial Stability Review 20*, 2016.

[2] E. W. T. Ngai, A. Gunasekaran, S. F. Wamba, S. Akter, and R. Dubey, "Big data analytics in electronic markets," *Electron. Mark.*, vol. 27, no. 3, pp. 243–245, 2017.

[3] L. Alexander, S. R. Das, Z. Ives, H. V. Jagadish, and C. Monteleoni, "Research Challenges in Financial Data Modeling and Analysis," 2017.

[4] J. J. J. M. Seddon and W. L. Currie, "A model for unpacking big data analytics in high-frequency trading," *J. Bus. Res.*, vol. 70, pp. 300–307, 2017.

[5] D. Laney, "Application Delivery Strategies." 2001.

[6] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Inf. Sci. (Ny).*, vol. 275, pp. 314–347, 2014.

[7] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Inf. Fusion*, vol. 28, pp. 45–59, 2016.

[8] I. Emmanuel and C. Stanier, "Defining Big Data," in *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies - BDAW '16*, 2016, pp. 1–6.

[9] Tom Shafer, "The 42 V's of Big Data and Data Science," 2017. [Online]. Available: https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html. [Accessed: 03-Nov-2017].

[10] J. Bleiholder and F. Naumann, "Data fusion," *ACM Comput. Surv.*, vol. 41, no. 1, pp. 1–41, 2008.

[11] M. M. Alyannezhadi, A. A. Pouyan, and V. Abolghasemi, "An efficient algorithm for multisensory data fusion under uncertainty condition," *J. Electr. Syst. Inf. Technol.*, vol. 4, no. 1, pp. 269–278, 2017.

[12] M. Välja, M. Korman, R. Lagerström, U. Franke, and M. Ekstedt, "Automated Architecture Modeling for Enterprise Technology Management Using Principles from Data Fusion : A Security Analysis Case," in *Proceedings of PICMET '16: Technology Management for Social Innovation*, 2016, pp. 14–22.

[13] V. Borges, "Survey of context information fusion for ubiquitous Internet-of-Things (IoT) systems," *Open Comput. Sci.*, vol. 6, no. 1, pp. 64–78, 2016.

[14] E. Blasch *et al.*, "Revisiting the JDL model for information exploitation," *Proc. 16th Int. Conf. Inf. Fusion, FUSION 2013*, pp. 129–136, 2013.

[15] D. Burdick *et al.*, "Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study.," *IEEE Data Eng. ...*, pp. 1–8, 2015.

[16] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proc. IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[17] F. Castanedo, "A review of data fusion techniques," *ScientificWorldJournal*, vol. 2013, 2013.

[18] E. Acar, M. A. Rasmussen, F. Savorani, T. Næs, and R. Bro, "Understanding data fusion within the framework of coupled matrix and tensor factorizations," *Chemom. Intell. Lab. Syst.*, vol. 129, pp. 53–63, 2013.

[19] K. Golmohammadi, O. R. Zaiane, S. Golmohammadi, K. Golmohammadi, and O. R. Zaiane, "Time series contextual anomaly detection for detecting market manipulation in stock market," 2015.

[20] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Inf. Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[21] L. C. Andersen, "Data-driven Approach to Information Sharing using Data Fusion and Machine Learning," Norwegian University of Science and Technology, 2016.

[22] T. Geva and J. Zahavi, "Predicting Intraday Stock Returns by Integrating Market Data and Financial News Reports Predicting Intraday Stock Returns by Integrating Market Data and Financial News Reports," in *Mediterranean Conference on Information Systems*, 2010.

[23] T. Geva and J. Zahavi, "Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news," *Decis. Support Syst.*, vol. 57, pp. 212–223, 2014.

[24] A. D. Mora, A. J. Falcão, L. Miranda, R. A. Ribeiro, and J. M. Fonseca, *Multisensor Data Fusion*. 2016.

[25] M. Bevilacqua, A. Tsourdos, A. Starr, and I. Durazo-Cardenas, "Data Fusion Strategy for Precise Vehicle Location for Intelligent Self-Aware Maintenance Systems," in *2015 6th International Conference on Intelligent Systems, Modelling and Simulation*, 2015.

[26] D. Mcdaniel, "An Information Fusion Framework for Data Integration," in *Information Fusion Application to Data Integration*, 2001, no. 858.

[27] A. Abdelgawad and M. Bayoumi, "Data Fusion in WSN," in *Resource-Aware Data Fusion Algorithms for Wireless Sensor Networks*, vol. 118, 2012.

[28] L. Snidaro, J. García, and J. Llinas, "Context-based Information Fusion: A survey and discussion," *Inf. Fusion*, vol. 25, pp. 16–31, 2015.

[29] B. Chandrasekaran, S. Gangadhar, and J. M. Conrad, "A survey of multisensor fusion techniques, architectures and methodologies," in *Conference Proceedings - IEEE SOUTHEASTCON*, 2017.

[30] J. R. (Jitendra R. . Raol, *Multi-sensor data fusion with MATLAB*. CRC Press, 2010.

[31] W. Elmenreich, "An Introduction to Sensor Fusion," 2002.

[32] M. A. Solano and G. Jernigan, "Enterprise data architecture principles for High-Level Multi-Int fusion: A pragmatic guide for implementing a heterogeneous data exploitation framework," in *Information Fusion (FUSION)*, 2012, pp. 867–874.

[33] A. N. Steinberg, C. L. Bowman, and F. E. White, "Revisions to the JDL data fusion model," *Proc. SPIE*, vol. 3719, no. 1, pp. 430–441, 1999.

[34] L. Wald, "Data fusion: A Conceptual Approach for an Efficient Exploitation of Remote Sensing Images," *Fusion Earth Data, Int. Conf.*, no. January, pp. 17–23, 1998.

[35] J. Llinas, C. Bowman, G. Rogova, and A. Steinberg, "Revisiting the JDL data fusion model II," *Sp. Nav. Warf. Syst. Command*, vol. 1, no. 7, pp. 1–14, 2004.

[36] Antweiler, Werner, Frank, and M. Z., "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," *J. Finance*, vol. 59, no. 3, pp. 1259–1294, 2004.

[37] H. M. Chen, "Group Polarization in Virtual Communities : The Case of Stock Message Boards," *Sch. Libr. Inf. Sci.*, no. 1994, pp. 185–195, 2013.

[38] F. Sun, A. Belatreche, S. Coleman, T. M. Mcginnity, and Y. Li, "Pre-processing Online Financial Text for Sentiment Classification: A Natural Language Processing Approach," in *Computational Intelligence for Financial Engineering & Economics (CIFEr)*, 2014.

[39] M. Owda, K. Crockett, and Pie Lee, "Financial Discussion Boards Irregularities Detection System (FDBs-IDS) using Information Extraction," in *Intelligent Systems*, 2017, no. September, pp. 8–12.

[40] M. Mirbabaie, S. Stieglitz, and M. Ruiz Eiro, "#IronyOff – Understanding the Usage of Irony on Twitter during a Corporate Crisis.," *Proc. Pacific Asia Conf. Inf. Syst. 2017*, no. July, 2017.

[41] M. Zappavigna, *The discourse of Twitter and social media*. Continuum International Pub. Group, 2012.

[42] L. Cazzoli, R. Sharma, M. Treccani, and F. Lillo, "A Large Scale Study to Understand the Relation between Twitter and Financial Market," in *2016 Third European Network Intelligence Conference (ENIC)*, 2016, pp. 98–105.

[43] A. Tafti, R. Zotti, and W. Jank, "Real-time diffusion of information on twitter and the financial markets," *PLoS One*, vol. 11, no. 8, pp. 1–16, 2016.

[44] A. F. Vilas, L. Evans, M. Owda, R. P. D. Redondo, and K. Crockett, "Experiment for analysing the impact of financial events on Twitter," in *Algorithms and Architectures for Parallel Processing*, 2017.

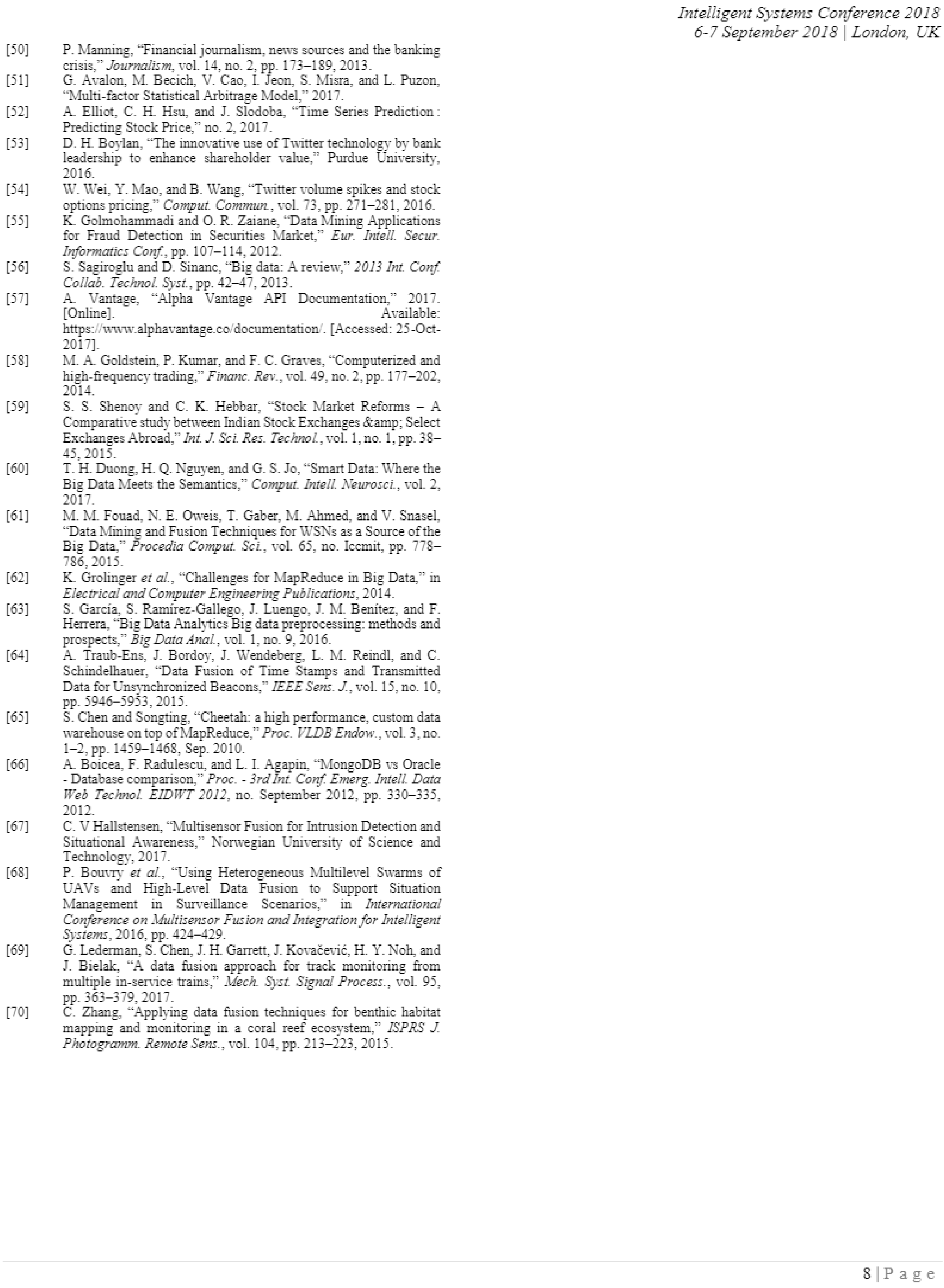
[45] H. Kwuan, "Twitter Cashtags and Sentiment Analysis in Predicting Stock Price Movements," 2017.

[46] A. Rosen, "Tweeting Made Easier," 2017. [Online]. Available: https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html. [Accessed: 08-Nov-2017].

[47] Q. Li, S. Shah, A. Nourbakhsh, R. Fang, and X. Liu, "funSentiment at SemEval-2017 Task 5: Fine-Grained Sentiment Anal- ysis on Financial Microblogs Using Word Vectors Built from StockTwits and Twitter," pp. 852–856, 2017.

[48] L. Harris, *Trading and Exchanges: Market Microstructure for Practitioners*, vol. 60, no. 4. Oxford University Press, 2002.

[49] London South East, "Broker Ratings," 2017. [Online]. Available: http://www.lse.co.uk/broker-tips.asp. [Accessed: 28-Oct-2017].



[50] P. Manning, "Financial journalism, news sources and the banking crisis," *Journalism*, vol. 14, no. 2, pp. 173–189, 2013.

[51] G. Avalon, M. Becich, V. Cao, I. Jeon, S. Misra, and L. Puzon, "Multi-factor Statistical Arbitrage Model," 2017.

[52] A. Elliot, C. H. Hsu, and J. Slodoba, "Time Series Prediction : Predicting Stock Price," no. 2, 2017.

[53] D. H. Boylan, "The innovative use of Twitter technology by bank leadership to enhance shareholder value," Purdue University, 2016.

[54] W. Wei, Y. Mao, and B. Wang, "Twitter volume spikes and stock options pricing," *Comput. Commun.*, vol. 73, pp. 271–281, 2016.

[55] K. Golmohammadi and O. R. Zaiane, "Data Mining Applications for Fraud Detection in Securities Market," *Eur. Intell. Secur. Informatics Conf.*, pp. 107–114, 2012.

[56] S. Sagiroglu and D. Sinanc, "Big data: A review," *2013 Int. Conf. Collab. Technol. Syst.*, pp. 42–47, 2013.

[57] A. Vantage, "Alpha Vantage API Documentation," 2017. [Online]. Available: https://www.alphavantage.co/documentation/. [Accessed: 25-Oct-2017].

[58] M. A. Goldstein, P. Kumar, and F. C. Graves, "Computerized and high-frequency trading," *Financ. Rev.*, vol. 49, no. 2, pp. 177–202, 2014.

[59] S. S. Shenoy and C. K. Hebbar, "Stock Market Reforms – A Comparative study between Indian Stock Exchanges &amp; Select Exchanges Abroad," *Int. J. Sci. Res. Technol.*, vol. 1, no. 1, pp. 38–45, 2015.

[60] T. H. Duong, H. Q. Nguyen, and G. S. Jo, "Smart Data: Where the Big Data Meets the Semantics," *Comput. Intell. Neurosci.*, vol. 2, 2017.

[61] M. M. Fouad, N. E. Oweis, T. Gaber, M. Ahmed, and V. Snasel, "Data Mining and Fusion Techniques for WSNs as a Source of the Big Data," *Procedia Comput. Sci.*, vol. 65, no. Iccmit, pp. 778–786, 2015.

[62] K. Grolinger *et al.*, "Challenges for MapReduce in Big Data," in *Electrical and Computer Engineering Publications*, 2014.

[63] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big Data Analytics Big data preprocessing: methods and prospects," *Big Data Anal.*, vol. 1, no. 9, 2016.

[64] A. Traub-Ens, J. Bordoy, J. Wendeberg, L. M. Reindl, and C. Schindelhauer, "Data Fusion of Time Stamps and Transmitted Data for Unsynchronized Beacons," *IEEE Sens. J.*, vol. 15, no. 10, pp. 5946–5953, 2015.

[65] S. Chen and Songting, "Cheetah: a high performance, custom data warehouse on top of MapReduce," *Proc. VLDB Endow.*, vol. 3, no. 1–2, pp. 1459–1468, Sep. 2010.

[66] A. Boicea, F. Radulescu, and L. I. Agapin, "MongoDB vs Oracle - Database comparison," *Proc. - 3rd Int. Conf. Emerg. Intell. Data Web Technol. EIDWT 2012*, no. September 2012, pp. 330–335, 2012.

[67] C. V Hallstensen, "Multisensor Fusion for Intrusion Detection and Situational Awareness," Norwegian University of Science and Technology, 2017.

[68] P. Bouvry *et al.*, "Using Heterogeneous Multilevel Swarms of UAVs and High-Level Data Fusion to Support Situation Management in Surveillance Scenarios," in *International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2016, pp. 424–429.

[69] G. Lederman, S. Chen, J. H. Garrett, J. Kovačević, H. Y. Noh, and J. Bielak, "A data fusion approach for track monitoring from multiple in-service trains," *Mech. Syst. Signal Process.*, vol. 95, pp. 363–379, 2017.

[70] C. Zhang, "Applying data fusion techniques for benthic habitat mapping and monitoring in a coral reef ecosystem," *ISPRS J. Photogramm. Remote Sens.*, vol. 104, pp. 213–223, 2015.

Appendix B

SDE Companies

TABLE B.1: SDE companies (Alternative Investment Market)

Company Ticker	Company Name	Company Industry
GGP	Greatland Gold Plc	Basic Materials
VRS	Versarien Plc	Basic Materials
KDNC	Cadence Minerals Plc	Basic Materials
BIOM	Biome Technologies Plc	Basic Materials
CRPR	Cropper (James) Plc	Basic Materials
PREM	Premier African Minerals Limited	Basic Materials
AAU	Ariana Resources Plc	Basic Materials
RRR	Red Rock Resources Plc	Basic Materials
HRN	Hornby Plc	Consumer Goods
MUL	Mulberry Group Plc	Consumer Goods
WYN	Wynnstay Group Plc	Consumer Goods
FEVR	Fevertree Drinks Plc	Consumer Goods
TUNE	Focusrite Plc	Consumer Goods
LWRF	Lightwaverf Plc	Consumer Goods
FDEV	Frontier Developments Plc	Consumer Goods
G4M	Gear4music (Holdings) Plc	Consumer Goods
HOTC	Hotel Chocolat Group Plc	Consumer Goods
SIS	Science In Sport Plc	Consumer Goods
TEF	Telford Homes Plc	Consumer Goods
ZAM	Zambeef Products Plc	Consumer Goods
ASC	Asos Plc	Consumer Services
EMAN	Everyman Media Group Plc	Consumer Services
JOUL	Joules Group Plc	Consumer Services
BOO	Boohoo.Com Plc	Consumer Services
KOOV	Koovs Plc	Consumer Services
YOU	Yougov Plc	Consumer Services
APGN	Applegreen Plc	Consumer Services
CCP	Celtic Plc	Consumer Services
CRAW	Crawshaw Group Plc	Consumer Services
FJET	Fastjet Plc	Consumer Services
SHOE	Shoe Zone Plc	Consumer Services
TMO	Time Out Group Plc	Consumer Services
UCG	United Carpets Group Plc	Consumer Services
HUNT	Hunters Property Plc	Financials
MTR	Metal Tiger Plc	Financials

CRC	Circle Property Plc	Financials
BLV	Belvoir Lettings Plc	Financials
TUNG	Tungsten Corporation Plc	Financials
PURP	Purplebricks Group Plc	Financials
ARGO	Argo Group Limited	Financials
MTW	Mattioli Woods Plc	Financials
TPFG	Property Franchise Group Plc (The)	Financials
PGH	Personal Group Holdings Plc	Financials
MAB1	Mortgage Advice Bureau (Holdings) Plc	Financials
ABC	Abcam Plc	Health Care
COG	Cambridge Cognition Holdings Plc	Health Care
AMYT	Amryt Pharma Plc	Health Care
CLIN	Clinigen Group Plc	Health Care
HZD	Horizon Discovery Group Plc	Health Care
AGL	Angle Plc	Health Care
AVCT	Avacta Group Plc	Health Care
KMK	Kromek Group Plc	Health Care
REDX	Redx Pharma Plc	Health Care
SUN	Surgical Innovations Group Plc	Health Care
SAR	Sareum Holdings Plc	Health Care
FLOW	Flowgroup Plc	Industrials
INSE	Inspired Energy Plc	Industrials
NAK	Nakama Group Plc	Industrials
DX.	Dx (Group) Plc	Industrials
WYG	Wyg Plc	Industrials
MRS	Management Resource Solutions Plc	Industrials
ASY	Andrews Sykes Group Plc	Industrials
BEG	Begbies Traynor Group Plc	Industrials
CTG	Christie Group Plc	Industrials
GTLY	Gateley (Holdings) Plc	Industrials
UTW	Utilitywise Plc	Industrials
88E	88 Energy Limited	Oil Gas
GBP	Global Petroleum Limited	Oil Gas
ITM	Itm Power Plc	Oil Gas
CLON	Clontarf Energy Plc	Oil Gas
NAUT	Nautilus Marine Services Plc	Oil Gas
SOU	Sound Energy Plc	Oil Gas
ANGS	Angus Energy Plc	Oil Gas
HUR	Hurricane Energy Plc	Oil Gas
NUOG	Nu-Oil And Gas Plc	Oil Gas
TLOU	Tlou Energy Limited	Oil Gas
SLE	San Leon Energy Plc	Oil Gas
EYE	Eagle Eye Solutions Group Plc	Technology
ING	Ingenta Plc	Technology
TRB	Tribal Group Plc	Technology
BGO	Bango Plc	Technology
WAND	Wandisco Plc	Technology
PRSM	Blue Prism Group Plc	Technology
ALB	Albert Technologies Ltd	Technology
AMO	Amino Technologies Plc	Technology

BBSN	Brave Bison Group Plc	Technology
ESG	Eservglobal Limited	Technology
FBT	Forbidden Technologies Plc	Technology
IOM	Iomart Group Plc	Technology
RDT	Rosslyn Data Technologies Plc	Technology
TCM	Telit Communications Plc	Technology
ZOO	Zoo Digital Group Plc	Technology
AVN	Avanti Communications Group Plc	Telecommunications
MANX	Manx Telecom Plc	Telecommunications
GAMA	Gamma Communications Plc	Telecommunications
MOS	Mobile Streams Plc	Telecommunications
TPOP	People's Operator Plc (The)	Telecommunications
GOOD	Good Energy Group Plc	Utilities
YU.	Yu Group Plc	Utilities
ACP	Armadale Capital Plc	Utilities

TABLE B.2: SDE companies (Main Market)

Company Ticker	**Company Name**	**Company Industry**
ACA	Acacia Mining Plc	Basic Materials
BFA	BASF Se	Basic Materials
BLT	BHP Billiton Plc	Basic Materials
PDL	Petra Diamonds Limited	Basic Materials
RIO	Rio Tinto Plc	Basic Materials
ZCC	ZCCM Investments Holdings Plc	Basic Materials
AAL	Anglo American Plc	Basic Materials
GLEN	Glencore Plc	Basic Materials
DGE	Diageo Plc	Consumer Goods
KNM	Konami Holdings Corporation	Consumer Goods
PSN	Persimmon Plc	Consumer Goods
TYT	Toyota Motor Corporation	Consumer Goods
BVIC	Britvic Plc	Consumer Goods
GAW	Games Workshop Group Plc	Consumer Goods
GNC	Greencore Group Plc	Consumer Goods
IMB	Imperial Brands Plc	Consumer Goods
RDW	Redrow Plc	Consumer Goods
ULVR	Unilever Plc	Consumer Goods
BMY	Bloomsbury Publishing Plc	Consumer Services
DEB	Debenhams Plc	Consumer Services
GMD	Game Digital Plc	Consumer Services
HFD	Halfords Group Plc	Consumer Services
MRW	Morrison (Wm) Supermarkets Plc	Consumer Services
TSCO	Tesco Plc	Consumer Services
AO.	AO World Plc	Consumer Services
CFYN	Caffyns Plc	Consumer Services
CCL	Carnival Plc	Consumer Services
CINE	Cineworld Group Plc	Consumer Services
FCCN	French Connection Group Plc	Consumer Services
MONY	Moneysupermarket.Com Group Plc	Consumer Services

PETS	Pets At Home Group Plc	Consumer Services
ADM	Admiral Group Plc	Financials
BARC	Barclays Plc	Financials
HSBA	HSBC Holdings Plc	Financials
SVS	Savills Plc	Financials
UAI	U And I Group Plc	Financials
RBS	Royal Bank Of Scotland Group Plc	Financials
ATMA	Atlas Mara Limited	Financials
BNC	Banco Santander S.A.	Financials
CAY	Charles Stanley Group Plc	Financials
GRI	Grainger Plc	Financials
MTRO	Metro Bank Plc	Financials
GNS	Genus Plc	Health Care
GSK	Glaxosmithkline Plc	Health Care
SHP	Shire Plc	Health Care
PRTC	Puretech Health Plc	Health Care
BTG	BTG Plc	Health Care
AZN	Astrazeneca Plc	Health Care
MDC	Mediclinic International Plc	Health Care
NMC	Nmc Health Plc	Health Care
DPH	Dechra Pharmaceuticals Plc	Health Care
SN.	Smith Nephew Plc	Health Care
HIK	Hikma Pharmaceuticals Plc	Health Care
BBYB	Balfour Beatty Plc	Industrials
ECM	Electrocomponents Plc	Industrials
GEC	General Electric Company	Industrials
KLR	Keller Group Plc	Industrials
RR.	Rolls-Royce Holdings Plc	Industrials
RMG	Royal Mail Plc	Industrials
AGK	Aggreko Plc	Industrials
CLLN	Carillion Plc	Industrials
ECEL	Eurocell Plc	Industrials
IMI	IMI Plc	Industrials
MTO	Mitie Group Plc	Industrials
BP.	BP Plc	Oil Gas
PMO	Premier Oil Plc	Oil Gas
TTA	Total S.A.	Oil Gas
WG.	Wood Group (John) Plc	Oil Gas
COPL	Canadian Overseas Petroleum Limited	Oil Gas
LKOH	PJSC Lukoil	Oil Gas
CNE	Cairn Energy Plc	Oil Gas
XPL	Xplorer Plc	Oil Gas
TLW	Tullow Oil Plc	Oil Gas
AVV	Aveva Group Plc	Technology
IBM	International Business Machines Corporation	Technology
SGE	Sage Group Plc	Technology
SDL	SDL Plc	Technology
SCT	Softcat Plc	Technology
USY	Unisys Corporation	Technology
CCC	Computacenter Plc	Technology

FDM	FDM Group (Holdings) Plc	Technology
NCC	NCC Group Plc	Technology
SOPH	Sophos Group Plc	Technology
TOOP	Toople Plc	Technology
KNOS	Kainos Group Plc	Technology
NANO	Nanoco Group Plc	Technology
RM.	RM Plc	Technology
SPT	Spirent Communications Plc	Technology
BT.A	BT Group Plc	Telecommunications
KCOM	KCOM Group Plc	Telecommunications
TDE	Telefonica Sa	Telecommunications
VOD	Vodafone Group Plc	Telecommunications
ISAT	Inmarsat Plc	Telecommunications
TALK	Talktalk Telecom Group Plc	Telecommunications
TEP	Telecom Plus	Telecommunications
CNA	Centrica Plc	Utilities
SVT	Severn Trent Plc	Utilities
UU.	United Utilities Group Plc	Utilities
DRX	Drax Group Plc	Utilities
PNN	Pennon Group Plc	Utilities

Appendix C

Cashtag Collision Paper

Expert Systems With Applications 127 (2019) 353–369

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A methodology for the resolution of cashtag collisions on Twitter – A natural language processing & data fusion approach

Lewis Evans[a,*], Majdi Owda[a], Keeley Crockett[a], Ana Fernandez Vilas[b]

[a] School of Computing, Mathematics & Digital Technology, Manchester Metropolitan University M1 5GD UK, Manchester, England United Kingdom
[b] I&C Lab. AtlantTIC Research Centre, University of Vigo, 36310, Pontevedra, Spain

ARTICLE INFO

Article history:
Received 22 November 2018
Revised 11 March 2019
Accepted 12 March 2019
Available online 12 March 2019

Keywords:
Cashtag collision
Twitter
Stock market
Data fusion
Machine learning
Natural language processing

ABSTRACT

Investors utilise social media such as Twitter as a means of sharing news surrounding financials stocks listed on international stock exchanges. Company ticker symbols are used to uniquely identify companies listed on stock exchanges and can be embedded within tweets to create clickable hyperlinks referred to as cashtags, allowing investors to associate their tweets with specific companies. The main limitation is that identical ticker symbols are present on exchanges all over the world, and when searching for such cashtags on Twitter, a stream of tweets is returned which match any company in which the cashtag refers to - we refer to this as a cashtag collision. The presence of colliding cashtags could sow confusion for investors seeking news regarding a specific company. A resolution to this issue would benefit investors who rely on the speediness of tweets for financial information, saving them precious time. We propose a methodology to resolve this problem which combines Natural Language Processing and Data Fusion to construct company-specific corpora to aid in the detection and resolution of colliding cashtags, so that tweets can be classified as being related to a specific stock exchange or not. Supervised machine learning classifiers are trained twice on each tweet – once on a count vectorisation of the tweet text, and again with the assistance of features contained in the company-specific corpora. We validate the cashtag collision methodology by carrying out an experiment involving companies listed on the London Stock Exchange. Results show that several machine learning classifiers benefit from the use of the custom corpora, yielding higher classification accuracy in the prediction and resolution of colliding cashtags.

1. Introduction

Investors make use of many online discussion channels when deciding to make investments on stock markets. Such information is presented within Financial Discussion Boards (FDBs), news corporations (e.g. Financial Times), broker agency websites, and social media platforms. Recently, Twitter has become a popular platform for investors to disseminate stock market information and discussion (Brown, 2012). Many large organisations are also using Twitter as a platform to obtain and share information relating to their products and services (Huizinga, Ayanso, Smoor, & Wronski, 2017).

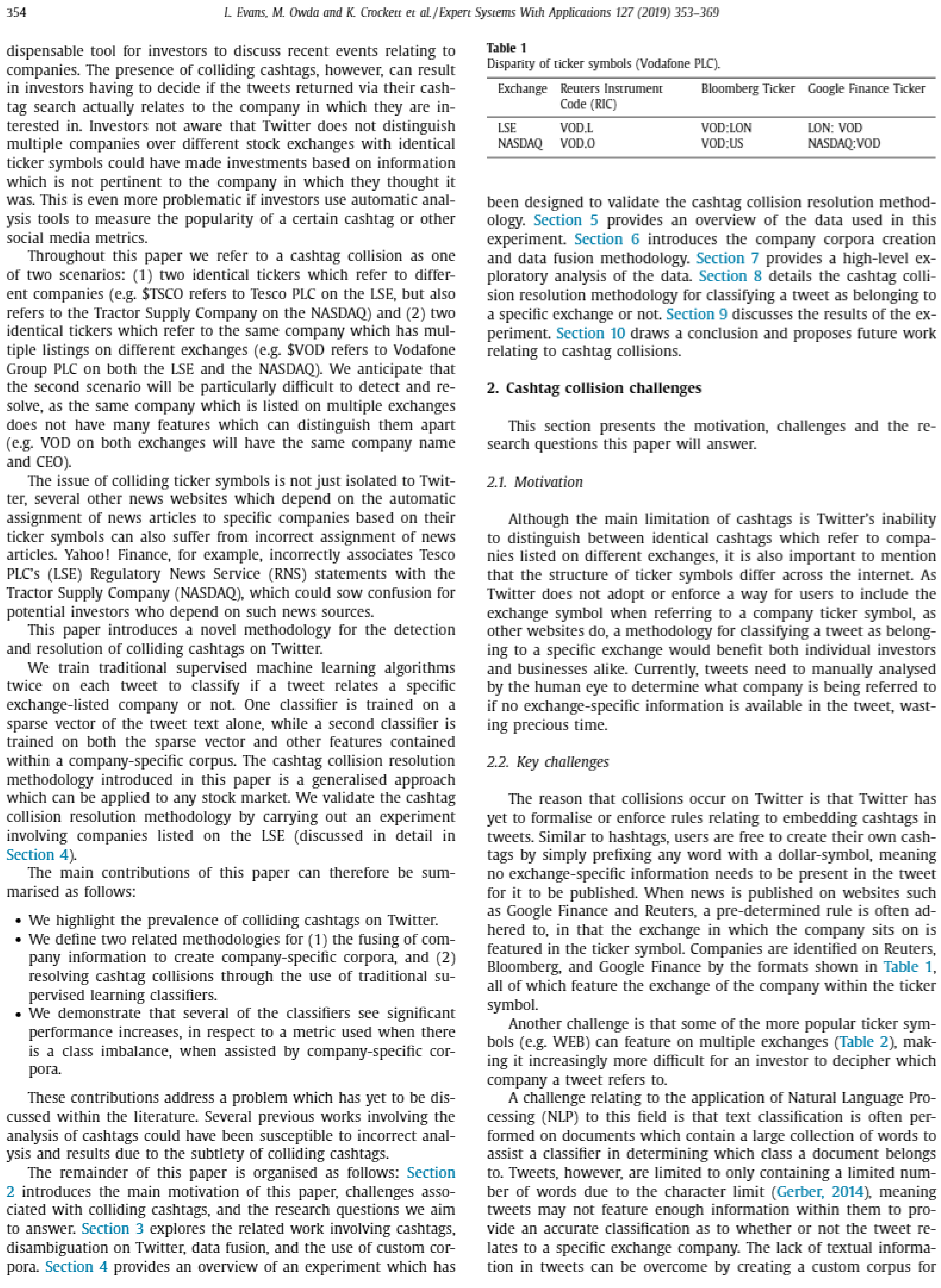
Companies are identified on stock markets through the use of ticker symbols, which are typically one to four characters in length (depending on the exchange) and are unique to an exchange, e.g. the TSCO ticker refers to Tesco PLC on the London Stock Exchange (LSE). The use of these ticker symbols within tweets on Twitter are referred to as cashtags and allow investors to participate in discussions and view news regarding a specific company at a moment's notice (Rajesh & Gandy, 2016). Cashtags are clickable links embedded within tweets which mimic the company's ticker symbol, prefixed with a dollar-symbol (e.g. $TSCO cashtag on Twitter refers to Tesco PLC) (Oliveira, Cortez, & Areal, 2016). Cashtags were originally introduced by Stocktwits[1] to allow users to link companies with their posts. Twitter introduced the feature of cashtags in 2012 to allow their users to associate specific companies with their tweets (Li, Shah, Nourbakhsh, Fang, & Liu, 2017). A tweet can contain multiple cashtags, with the only limitation being the character limit imposed upon Tweets, which was recently increased to 280 characters.

The main limitation of cashtags is that they are susceptible to colliding with an identical cashtag belonging to a company listed on another exchange, a phenomenon we refer to as a cashtag collision. As tweets are typically short in length, they can be an in-

[*] Corresponding author.
E-mail addresses: l.evans@mmu.ac.uk (L. Evans), m.owda@mmu.ac.uk (M. Owda), k.crockett@mmu.ac.uk (K. Crockett), avilas@det.uvigo.es (A.F. Vilas).

[1] https://stocktwits.com/.

https://doi.org/10.1016/j.eswa.2019.03.019



dispensable tool for investors to discuss recent events relating to companies. The presence of colliding cashtags, however, can result in investors having to decide if the tweets returned via their cashtag search actually relates to the company in which they are interested in. Investors not aware that Twitter does not distinguish multiple companies over different stock exchanges with identical ticker symbols could have made investments based on information which is not pertinent to the company in which they thought it was. This is even more problematic if investors use automatic analysis tools to measure the popularity of a certain cashtag or other social media metrics.

Throughout this paper we refer to a cashtag collision as one of two scenarios: (1) two identical tickers which refer to different companies (e.g. $TSCO refers to Tesco PLC on the LSE, but also refers to the Tractor Supply Company on the NASDAQ) and (2) two identical tickers which refer to the same company which has multiple listings on different exchanges (e.g. $VOD refers to Vodafone Group PLC on both the LSE and the NASDAQ). We anticipate that the second scenario will be particularly difficult to detect and resolve, as the same company which is listed on multiple exchanges does not have many features which can distinguish them apart (e.g. VOD on both exchanges will have the same company name and CEO).

The issue of colliding ticker symbols is not just isolated to Twitter, several other news websites which depend on the automatic assignment of news articles to specific companies based on their ticker symbols can also suffer from incorrect assignment of news articles. Yahoo! Finance, for example, incorrectly associates Tesco PLC's (LSE) Regulatory News Service (RNS) statements with the Tractor Supply Company (NASDAQ), which could sow confusion for potential investors who depend on such news sources.

This paper introduces a novel methodology for the detection and resolution of colliding cashtags on Twitter.

We train traditional supervised machine learning algorithms twice on each tweet to classify if a tweet relates a specific exchange-listed company or not. One classifier is trained on a sparse vector of the tweet text alone, while a second classifier is trained on both the sparse vector and other features contained within a company-specific corpus. The cashtag collision resolution methodology introduced in this paper is a generalised approach which can be applied to any stock market. We validate the cashtag collision resolution methodology by carrying out an experiment involving companies listed on the LSE (discussed in detail in Section 4).

The main contributions of this paper can therefore be summarised as follows:

- We highlight the prevalence of colliding cashtags on Twitter.
- We define two related methodologies for (1) the fusing of company information to create company-specific corpora, and (2) resolving cashtag collisions through the use of traditional supervised learning classifiers.
- We demonstrate that several of the classifiers see significant performance increases, in respect to a metric used when there is a class imbalance, when assisted by company-specific corpora.

These contributions address a problem which has yet to be discussed within the literature. Several previous works involving the analysis of cashtags could have been susceptible to incorrect analysis and results due to the subtlety of colliding cashtags.

The remainder of this paper is organised as follows: Section 2 introduces the main motivation of this paper, challenges associated with colliding cashtags, and the research questions we aim to answer. Section 3 explores the related work involving cashtags, disambiguation on Twitter, data fusion, and the use of custom corpora. Section 4 provides an overview of an experiment which has been designed to validate the cashtag collision resolution methodology. Section 5 provides an overview of the data used in this experiment. Section 6 introduces the company corpora creation and data fusion methodology. Section 7 provides a high-level exploratory analysis of the data. Section 8 details the cashtag collision resolution methodology for classifying a tweet as belonging to a specific exchange or not. Section 9 discusses the results of the experiment. Section 10 draws a conclusion and proposes future work relating to cashtag collisions.

Table 1
Disparity of ticker symbols (Vodafone PLC).

Exchange	Reuters Instrument Code (RIC)	Bloomberg Ticker	Google Finance Ticker
LSE	VOD.L	VOD:LON	LON: VOD
NASDAQ	VOD.O	VOD:US	NASDAQ:VOD

2. Cashtag collision challenges

This section presents the motivation, challenges and the research questions this paper will answer.

2.1. Motivation

Although the main limitation of cashtags is Twitter's inability to distinguish between identical cashtags which refer to companies listed on different exchanges, it is also important to mention that the structure of ticker symbols differ across the internet. As Twitter does not adopt or enforce a way for users to include the exchange symbol when referring to a company ticker symbol, as other websites do, a methodology for classifying a tweet as belonging to a specific exchange would benefit both individual investors and businesses alike. Currently, tweets need to manually analysed by the human eye to determine what company is being referred to if no exchange-specific information is available in the tweet, wasting precious time.

2.2. Key challenges

The reason that collisions occur on Twitter is that Twitter has yet to formalise or enforce rules relating to embedding cashtags in tweets. Similar to hashtags, users are free to create their own cashtags by simply prefixing any word with a dollar-symbol, meaning no exchange-specific information needs to be present in the tweet for it to be published. When news is published on websites such as Google Finance and Reuters, a pre-determined rule is often adhered to, in that the exchange in which the company sits on is featured in the ticker symbol. Companies are identified on Reuters, Bloomberg, and Google Finance by the formats shown in Table 1, all of which feature the exchange of the company within the ticker symbol.

Another challenge is that some of the more popular ticker symbols (e.g. WEB) can feature on multiple exchanges (Table 2), making it increasingly more difficult for an investor to decipher which company a tweet refers to.

A challenge relating to the application of Natural Language Processing (NLP) to this field is that text classification is often performed on documents which contain a large collection of words to assist a classifier in determining which class a document belongs to. Tweets, however, are limited to only containing a limited number of words due to the character limit (Gerber, 2014), meaning tweets may not feature enough information within them to provide an accurate classification as to whether or not the tweet relates to a specific exchange company. The lack of textual information in tweets can be overcome by creating a custom corpus for

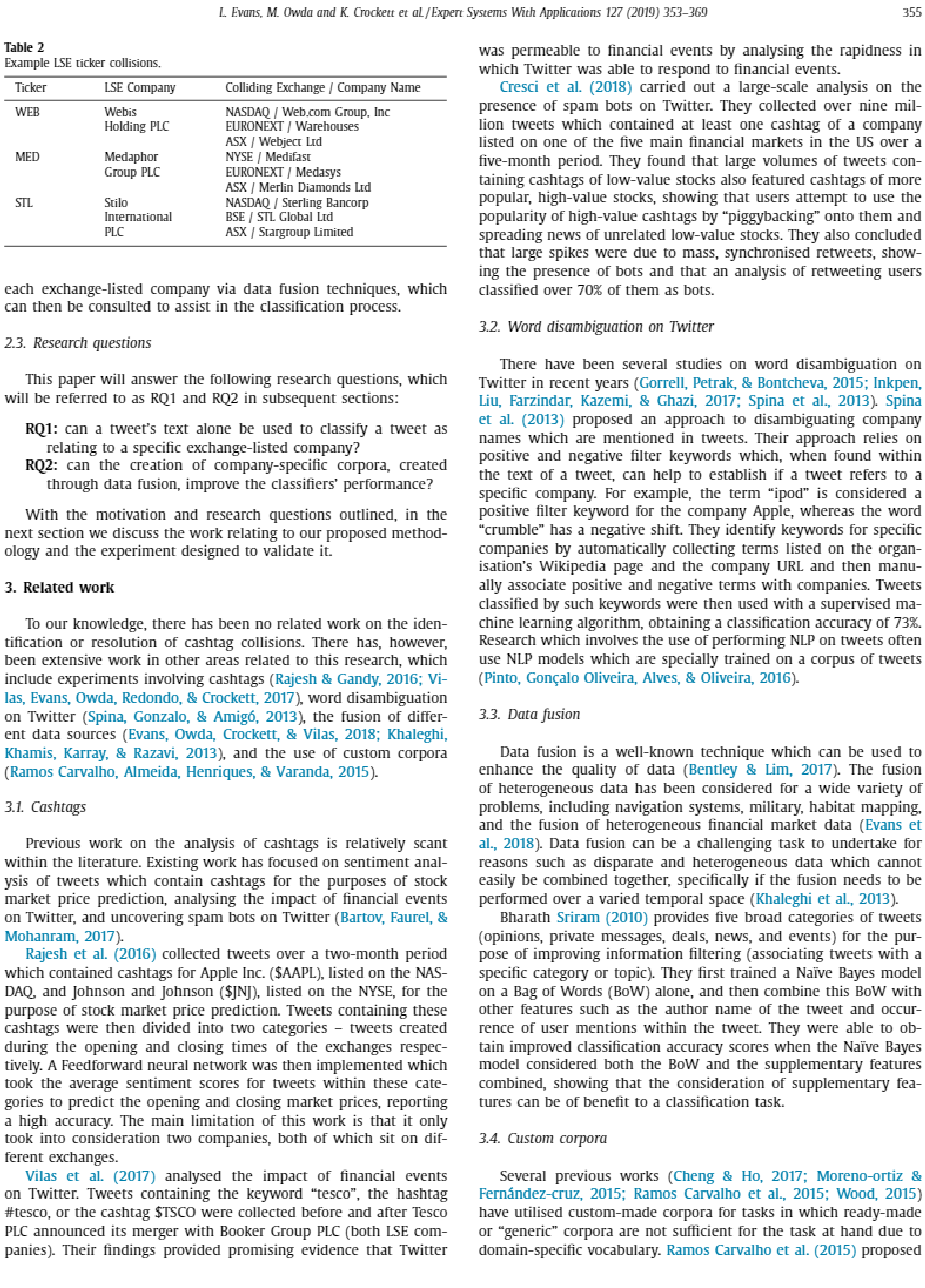


Table 2
Example LSE ticker collisions.

Ticker	LSE Company	Colliding Exchange / Company Name
WEB	Webis Holding PLC	NASDAQ / Web.com Group, Inc EURONEXT / Warehouses ASX / Webject Ltd
MED	Medaphor Group PLC	NYSE / Medifast EURONEXT / Medasys ASX / Merlin Diamonds Ltd
STL	Stilo International PLC	NASDAQ / Sterling Bancorp BSE / STL Global Ltd ASX / Stargroup Limited

each exchange-listed company via data fusion techniques, which can then be consulted to assist in the classification process.

2.3. Research questions

This paper will answer the following research questions, which will be referred to as RQ1 and RQ2 in subsequent sections:

RQ1: can a tweet's text alone be used to classify a tweet as relating to a specific exchange-listed company?
RQ2: can the creation of company-specific corpora, created through data fusion, improve the classifiers' performance?

With the motivation and research questions outlined, in the next section we discuss the work relating to our proposed methodology and the experiment designed to validate it.

3. Related work

To our knowledge, there has been no related work on the identification or resolution of cashtag collisions. There has, however, been extensive work in other areas related to this research, which include experiments involving cashtags (Rajesh & Gandy, 2016; Vilas, Evans, Owda, Redondo, & Crockett, 2017), word disambiguation on Twitter (Spina, Gonzalo, & Amigó, 2013), the fusion of different data sources (Evans, Owda, Crockett, & Vilas, 2018; Khaleghi, Khamis, Karray, & Razavi, 2013), and the use of custom corpora (Ramos Carvalho, Almeida, Henriques, & Varanda, 2015).

3.1. Cashtags

Previous work on the analysis of cashtags is relatively scant within the literature. Existing work has focused on sentiment analysis of tweets which contain cashtags for the purposes of stock market price prediction, analysing the impact of financial events on Twitter, and uncovering spam bots on Twitter (Bartov, Faurel, & Mohanram, 2017).

Rajesh et al. (2016) collected tweets over a two-month period which contained cashtags for Apple Inc. ($AAPL), listed on the NASDAQ, and Johnson and Johnson ($JNJ), listed on the NYSE, for the purpose of stock market price prediction. Tweets containing these cashtags were then divided into two categories – tweets created during the opening and closing times of the exchanges respectively. A Feedforward neural network was then implemented which took the average sentiment scores for tweets within these categories to predict the opening and closing market prices, reporting a high accuracy. The main limitation of this work is that it only took into consideration two companies, both of which sit on different exchanges.

Vilas et al. (2017) analysed the impact of financial events on Twitter. Tweets containing the keyword “tesco”, the hashtag #tesco, or the cashtag $TSCO were collected before and after Tesco PLC announced its merger with Booker Group PLC (both LSE companies). Their findings provided promising evidence that Twitter was permeable to financial events by analysing the rapidness in which Twitter was able to respond to financial events.

Cresci et al. (2018) carried out a large-scale analysis on the presence of spam bots on Twitter. They collected over nine million tweets which contained at least one cashtag of a company listed on one of the five main financial markets in the US over a five-month period. They found that large volumes of tweets containing cashtags of low-value stocks also featured cashtags of more popular, high-value stocks, showing that users attempt to use the popularity of high-value cashtags by “piggybacking” onto them and spreading news of unrelated low-value stocks. They also concluded that large spikes were due to mass, synchronised retweets, showing the presence of bots and that an analysis of retweeting users classified over 70% of them as bots.

3.2. Word disambiguation on Twitter

There have been several studies on word disambiguation on Twitter in recent years (Gorrell, Petrak, & Bontcheva, 2015; Inkpen, Liu, Farzindar, Kazemi, & Ghazi, 2017; Spina et al., 2013). Spina et al. (2013) proposed an approach to disambiguating company names which are mentioned in tweets. Their approach relies on positive and negative filter keywords which, when found within the text of a tweet, can help to establish if a tweet refers to a specific company. For example, the term “ipod” is considered a positive filter keyword for the company Apple, whereas the word “crumble” has a negative shift. They identify keywords for specific companies by automatically collecting terms listed on the organisation's Wikipedia page and the company URL and then manually associate positive and negative terms with companies. Tweets classified by such keywords were then used with a supervised machine learning algorithm, obtaining a classification accuracy of 73%. Research which involves the use of performing NLP on tweets often use NLP models which are specially trained on a corpus of tweets (Pinto, Gonçalo Oliveira, Alves, & Oliveira, 2016).

3.3. Data fusion

Data fusion is a well-known technique which can be used to enhance the quality of data (Bentley & Lim, 2017). The fusion of heterogeneous data has been considered for a wide variety of problems, including navigation systems, military, habitat mapping, and the fusion of heterogeneous financial market data (Evans et al., 2018). Data fusion can be a challenging task to undertake for reasons such as disparate and heterogeneous data which cannot easily be combined together, specifically if the fusion needs to be performed over a varied temporal space (Khaleghi et al., 2013).

Bharath Sriram (2010) provides five broad categories of tweets (opinions, private messages, deals, news, and events) for the purpose of improving information filtering (associating tweets with a specific category or topic). They first trained a Naïve Bayes model on a Bag of Words (BoW) alone, and then combine this BoW with other features such as the author name of the tweet and occurrence of user mentions within the tweet. They were able to obtain improved classification accuracy scores when the Naïve Bayes model considered both the BoW and the supplementary features combined, showing that the consideration of supplementary features can be of benefit to a classification task.

3.4. Custom corpora

Several previous works (Cheng & Ho, 2017; Moreno-ortiz & Fernández-cruz, 2015; Ramos Carvalho et al., 2015; Wood, 2015) have utilised custom-made corpora for tasks in which ready-made or “generic” corpora are not sufficient for the task at hand due to domain-specific vocabulary. Ramos Carvalho et al. (2015) proposed

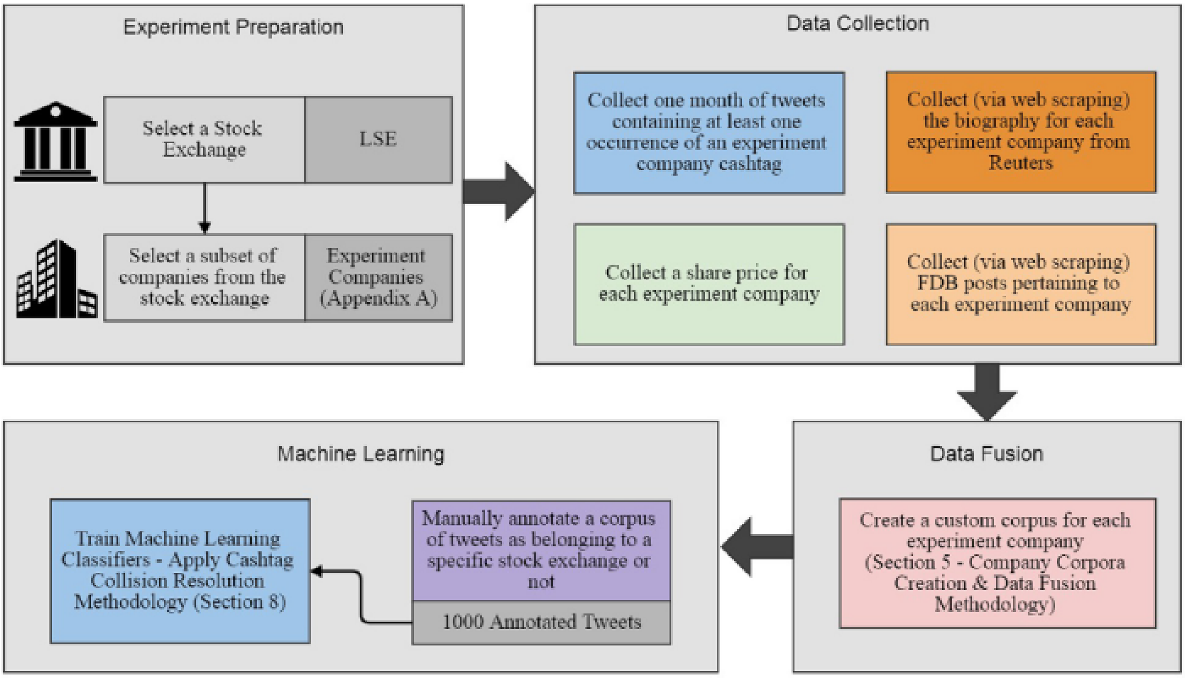

Experiment Preparation
Select a Stock Exchange
LSE
Select a subset of companies from the stock exchange
Experiment Companies (Appendix A)
Data Collection
Collect one month of tweets containing at least one occurrence of an experiment company cashtag
Collect (via web scraping) the biography for each experiment company from Reuters
Collect a share price for each experiment company
Collect (via web scraping) FDB posts pertaining to each experiment company
Data Fusion
Create a custom corpus for each experiment company (Section 5 - Company Corpora Creation & Data Fusion Methodology)
Machine Learning
Manually annotate a corpus of tweets as belonging to a specific stock exchange or not
1000 Annotated Tweets
Train Machine Learning Classifiers - Apply Cashtag Collision Resolution Methodology (Section 8)

Fig. 1. Experiment overview.

a technique to create domain-specific corpora to convert source code identifiers to their equivalent full name counterparts (e.g. a method named “strcmp” can be split into the words “string, compare”). Their work did note limitations in that, without a domain corpus, translations between source code identifiers to full words can be difficult to achieve.

This paper attempts to address several of the challenges outlined in the related work we have just explored. In regards to cashtag analysis, we consider a larger cashtag space than that explored in (Rajesh & Gandy, 2016) by examining 100 company cashtags. Although we do not attempt to disambiguate between specific keywords found within tweets, we do attempt to disambiguate tweets by classifying tweets as relating to an exchange-listed company or not. In regard to data fusion, we do not attempt to fuse data based on time. Instead, we fuse company-specific information together from three different external data sources in one batch, eliminating the challenges associating with real-time data fusion. This fusion process supports the creation of custom company corpora which will contain information that is specific to each company.

The next section will provide a high-level overview of an experiment to validate the cashtag collision resolution methodology.

4. Experiment details

An experiment (Fig. 1) has been designed which involves creating a custom corpus of company-specific information for 100 pre-selected companies.

4.1. Experiment preparation

For the purposes of this paper, we validate our cashtag collision resolution methodology by performing an experiment using 100 LSE companies (listed in Appendix A). The LSE has been chosen due to having a popular FDB associated with it which is dedicated to LSE-listed companies, allowing web scraping techniques to yield information specific to companies listed on that exchange. The LSE is formed of two sub-markets; the Alternative Investment Market (AIM) and the Main Market (MM). The AIM is suited for growing businesses and has a more flexible regulatory system than the MM (Barnes, 2017).

4.2. Company selection

In regards to the 100 companies used in our experiment, we select 50 companies from each sub-market (25 of which have a known collision with another company listed on one of the exchanges in Table 3, the remaining 25 with no known collision with the exchanges). Companies are selected randomly from each of the LSE’s ten different industries (basic materials, consumer goods, consumer services, financials, health care, industrials, oil & gas, technology, telecommunications, and utilities). Only companies which have been listed on the LSE for at least two years were eligible in this selection process, to ensure that they are well-established and to maximise the chance of collecting tweets containing cashtags relating to LSE-listed companies.

4.2.1. Data collection

In order to ascertain if a tweet relates to a specific exchange-listed company, such as the LSE, data from multiple, reputable sources will be collected and combined to ensure a reliable reference to each of the LSE-listed companies is available.

Tweets pertaining to the 100 experiment companies are collected in real-time via the Twitter Streaming API, which collects no more than 1% of all tweets tweeted in real-time (Abdeen, Wu, Erickson, & Fandy, 2015). Descriptions for each of these companies are web scraped from Reuters so that certain keywords associated with the LSE-listed cashtag company can be obtained, which will be beneficial later to ascertain how many words within the tweets are also found to be in LSE-listed company’s biography. FDB posts are then collected from an FDB which is dedicated to LSE companies, allowing us to collect posts which are specific to the LSE companies used in this experiment.

Finally, a share price for the company is collected to assist in the manual annotation of the tweets, this can be a helpful attribute

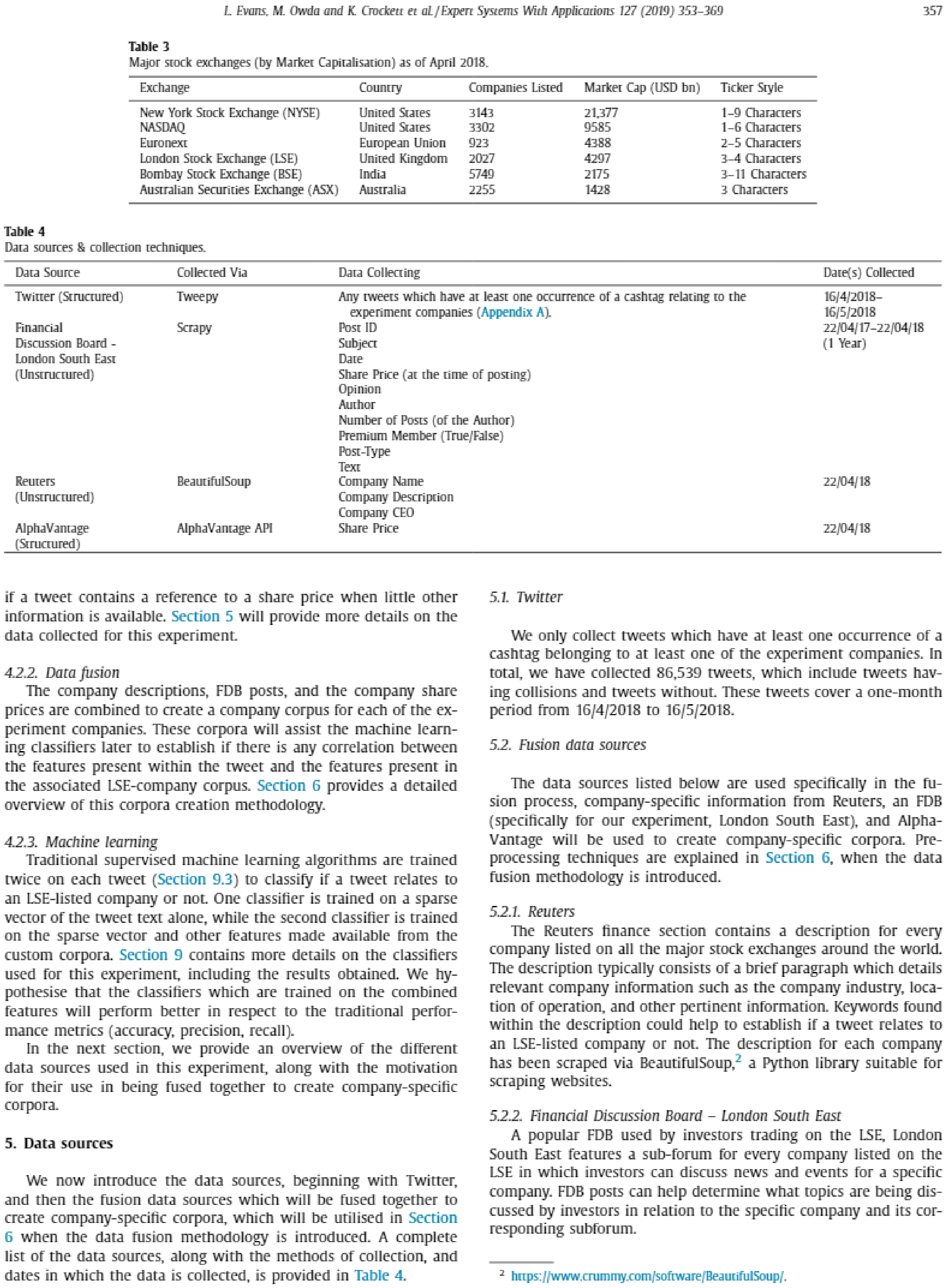


Table 3
Major stock exchanges (by Market Capitalisation) as of April 2018.

Exchange	Country	Companies Listed	Market Cap (USD bn)	Ticker Style
New York Stock Exchange (NYSE)	United States	3143	21,377	1–9 Characters
NASDAQ	United States	3302	9585	1–6 Characters
Euronext	European Union	923	4388	2–5 Characters
London Stock Exchange (LSE)	United Kingdom	2027	4297	3–4 Characters
Bombay Stock Exchange (BSE)	India	5749	2175	3–11 Characters
Australian Securities Exchange (ASX)	Australia	2255	1428	3 Characters

Table 4
Data sources & collection techniques.

Data Source	Collected Via	Data Collecting	Date(s) Collected
Twitter (Structured)	Tweepy	Any tweets which have at least one occurrence of a cashtag relating to the experiment companies (Appendix A).	16/4/2018–16/5/2018
Financial Discussion Board - London South East (Unstructured)	Scrapy	Post ID Subject Date Share Price (at the time of posting) Opinion Author Number of Posts (of the Author) Premium Member (True/False) Post-Type Text	22/04/17–22/04/18 (1 Year)
Reuters (Unstructured)	BeautifulSoup	Company Name Company Description Company CEO	22/04/18
AlphaVantage (Structured)	AlphaVantage API	Share Price	22/04/18

if a tweet contains a reference to a share price when little other information is available. Section 5 will provide more details on the data collected for this experiment.

4.2.2. Data fusion

The company descriptions, FDB posts, and the company share prices are combined to create a company corpus for each of the experiment companies. These corpora will assist the machine learning classifiers later to establish if there is any correlation between the features present within the tweet and the features present in the associated LSE-company corpus. Section 6 provides a detailed overview of this corpora creation methodology.

4.2.3. Machine learning

Traditional supervised machine learning algorithms are trained twice on each tweet (Section 9.3) to classify if a tweet relates to an LSE-listed company or not. One classifier is trained on a sparse vector of the tweet text alone, while the second classifier is trained on the sparse vector and other features made available from the custom corpora. Section 9 contains more details on the classifiers used for this experiment, including the results obtained. We hypothesise that the classifiers which are trained on the combined features will perform better in respect to the traditional performance metrics (accuracy, precision, recall).

In the next section, we provide an overview of the different data sources used in this experiment, along with the motivation for their use in being fused together to create company-specific corpora.

5. Data sources

We now introduce the data sources, beginning with Twitter, and then the fusion data sources which will be fused together to create company-specific corpora, which will be utilised in Section 6 when the data fusion methodology is introduced. A complete list of the data sources, along with the methods of collection, and dates in which the data is collected, is provided in Table 4.

5.1. Twitter

We only collect tweets which have at least one occurrence of a cashtag belonging to at least one of the experiment companies. In total, we have collected 86,539 tweets, which include tweets having collisions and tweets without. These tweets cover a one-month period from 16/4/2018 to 16/5/2018.

5.2. Fusion data sources

The data sources listed below are used specifically in the fusion process, company-specific information from Reuters, an FDB (specifically for our experiment, London South East), and AlphaVantage will be used to create company-specific corpora. Preprocessing techniques are explained in Section 6, when the data fusion methodology is introduced.

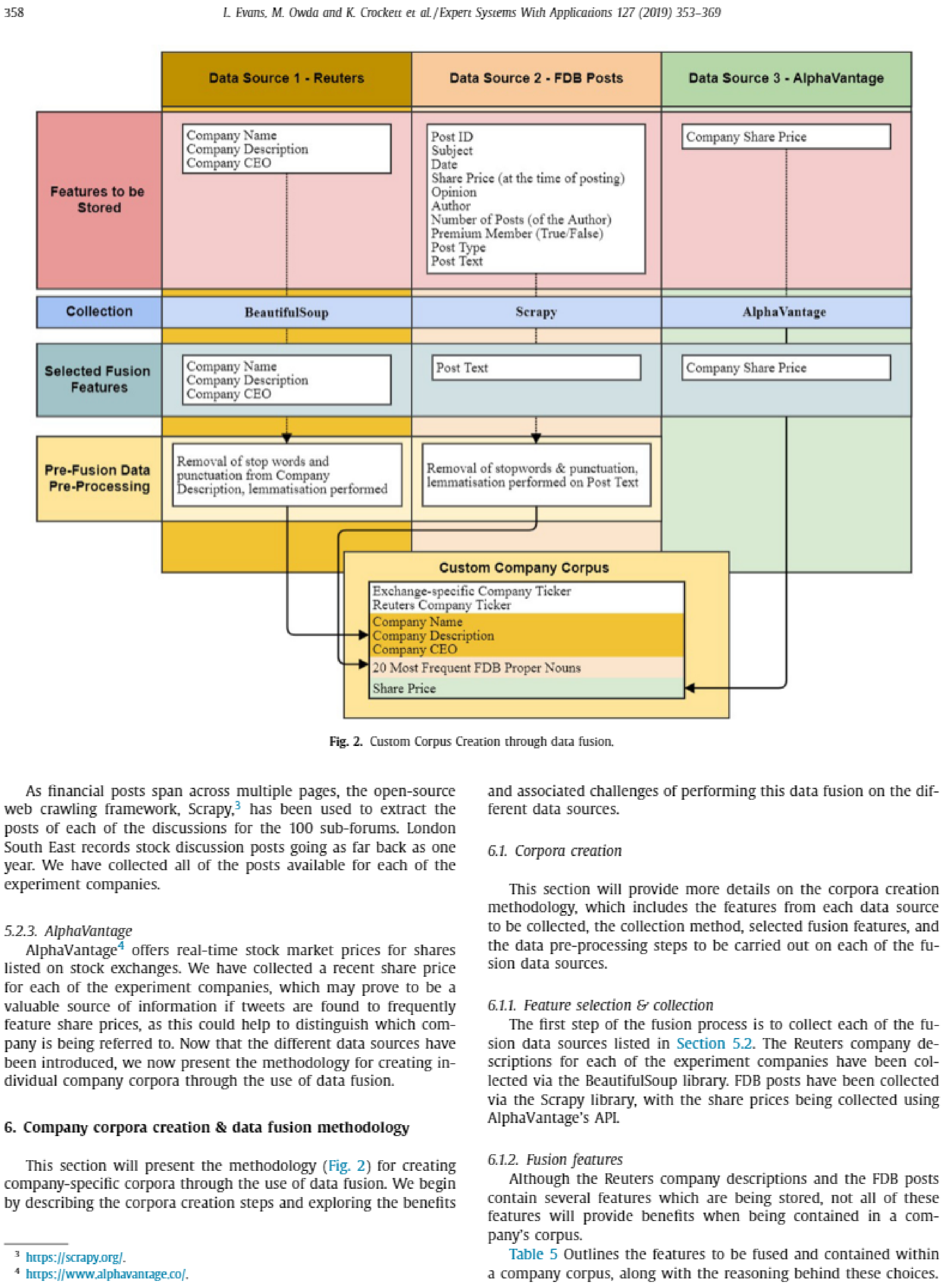
5.2.1. Reuters

The Reuters finance section contains a description for every company listed on all the major stock exchanges around the world. The description typically consists of a brief paragraph which details relevant company information such as the company industry, location of operation, and other pertinent information. Keywords found within the description could help to establish if a tweet relates to an LSE-listed company or not. The description for each company has been scraped via BeautifulSoup,[2] a Python library suitable for scraping websites.

5.2.2. Financial Discussion Board – London South East

A popular FDB used by investors trading on the LSE, London South East features a sub-forum for every company listed on the LSE in which investors can discuss news and events for a specific company. FDB posts can help determine what topics are being discussed by investors in relation to the specific company and its corresponding subforum.

[2] https://www.crummy.com/software/BeautifulSoup/.



Data Source 1 - Reuters
Data Source 2 - FDB Posts
Data Source 3 - AlphaVantage
Features to be Stored
Company Name
Company Description
Company CEO
Post ID
Subject
Date
Share Price (at the time of posting)
Opinion
Author
Number of Posts (of the Author)
Premium Member (True/False)
Post Type
Post Text
Company Share Price
Collection
BeautifulSoup
Scrapy
AlphaVantage
Selected Fusion Features
Company Name
Company Description
Company CEO
Post Text
Company Share Price
Pre-Fusion Data Pre-Processing
Removal of stop words and punctuation from Company Description, lemmatisation performed
Removal of stopwords & punctuation, lemmatisation performed on Post Text
Custom Company Corpus
Exchange-specific Company Ticker
Reuters Company Ticker
Company Name
Company Description
Company CEO
20 Most Frequent FDB Proper Nouns
Share Price

Fig. 2. Custom Corpus Creation through data fusion.

As financial posts span across multiple pages, the open-source web crawling framework, Scrapy,[3] has been used to extract the posts of each of the discussions for the 100 sub-forums. London South East records stock discussion posts going as far back as one year. We have collected all of the posts available for each of the experiment companies.

5.2.3. AlphaVantage

AlphaVantage[4] offers real-time stock market prices for shares listed on stock exchanges. We have collected a recent share price for each of the experiment companies, which may prove to be a valuable source of information if tweets are found to frequently feature share prices, as this could help to distinguish which company is being referred to. Now that the different data sources have been introduced, we now present the methodology for creating individual company corpora through the use of data fusion.

6. Company corpora creation & data fusion methodology

This section will present the methodology (Fig. 2) for creating company-specific corpora through the use of data fusion. We begin by describing the corpora creation steps and exploring the benefits and associated challenges of performing this data fusion on the different data sources.

6.1. Corpora creation

This section will provide more details on the corpora creation methodology, which includes the features from each data source to be collected, the collection method, selected fusion features, and the data pre-processing steps to be carried out on each of the fusion data sources.

6.1.1. Feature selection & collection

The first step of the fusion process is to collect each of the fusion data sources listed in Section 5.2. The Reuters company descriptions for each of the experiment companies have been collected via the BeautifulSoup library. FDB posts have been collected via the Scrapy library, with the share prices being collected using AlphaVantage's API.

6.1.2. Fusion features

Although the Reuters company descriptions and the FDB posts contain several features which are being stored, not all of these features will provide benefits when being contained in a company's corpus.

Table 5 Outlines the features to be fused and contained within a company corpus, along with the reasoning behind these choices.

[3] https://scrapy.org/.

[4] https://www.alphavantage.co/.

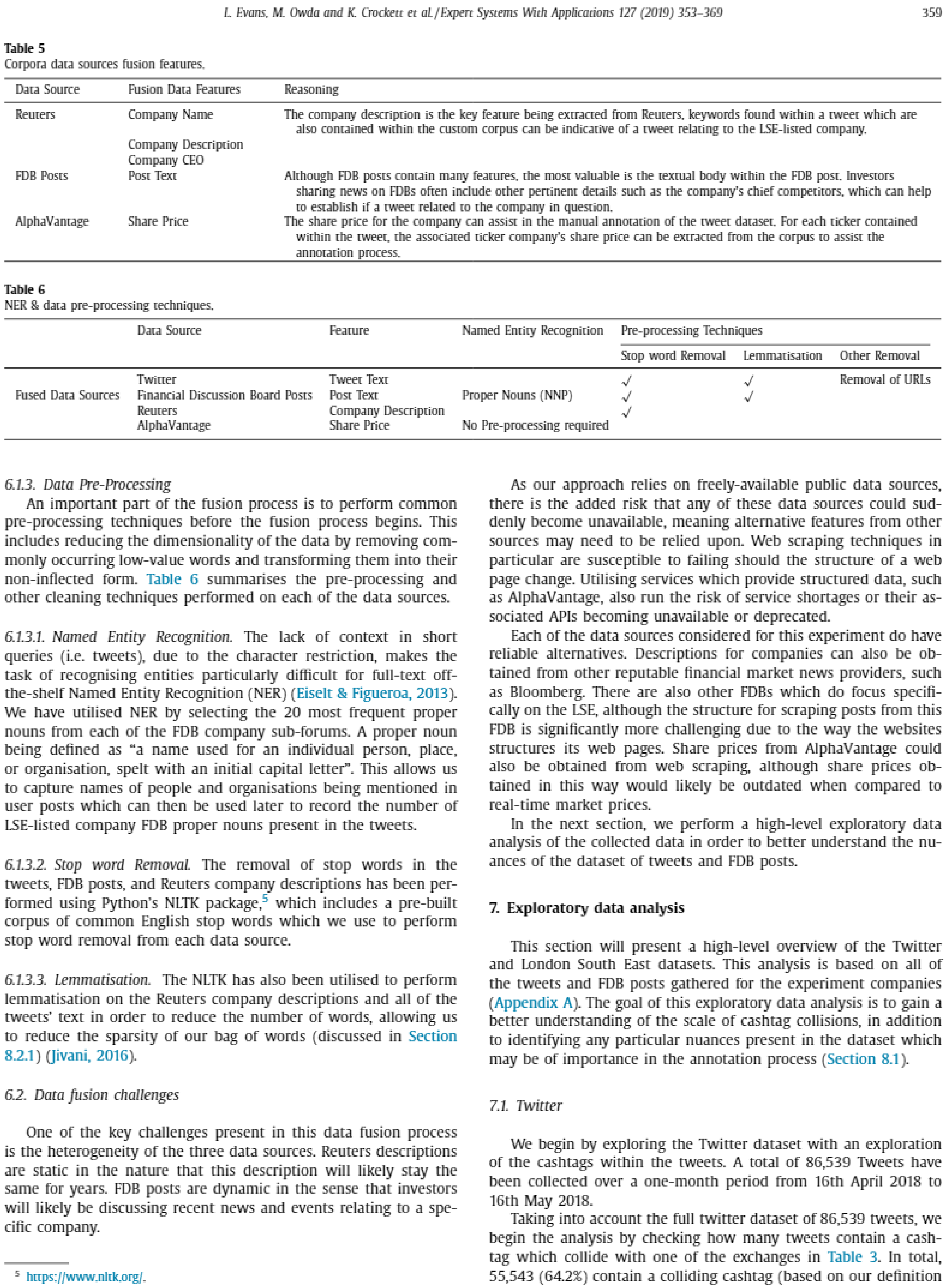


Table 5
Corpora data sources fusion features.

Data Source	Fusion Data Features	Reasoning
Reuters	Company Name	The company description is the key feature being extracted from Reuters, keywords found within a tweet which are also contained within the custom corpus can be indicative of a tweet relating to the LSE-listed company.
	Company Description	
	Company CEO	
FDB Posts	Post Text	Although FDB posts contain many features, the most valuable is the textual body within the FDB post. Investors sharing news on FDBs often include other pertinent details such as the company's chief competitors, which can help to establish if a tweet related to the company in question.
AlphaVantage	Share Price	The share price for the company can assist in the manual annotation of the tweet dataset. For each ticker contained within the tweet, the associated ticker company's share price can be extracted from the corpus to assist the annotation process.

Table 6
NER & data pre-processing techniques.

	Data Source	Feature	Named Entity Recognition	Pre-processing Techniques		
				Stop word Removal	Lemmatisation	Other Removal
	Twitter	Tweet Text		✓	✓	Removal of URLs
Fused Data Sources	Financial Discussion Board Posts	Post Text	Proper Nouns (NNP)	✓	✓	
	Reuters	Company Description		✓		
	AlphaVantage	Share Price	No Pre-processing required			

6.1.3. Data Pre-Processing

An important part of the fusion process is to perform common pre-processing techniques before the fusion process begins. This includes reducing the dimensionality of the data by removing commonly occurring low-value words and transforming them into their non-inflected form. Table 6 summarises the pre-processing and other cleaning techniques performed on each of the data sources.

6.1.3.1. Named Entity Recognition. The lack of context in short queries (i.e. tweets), due to the character restriction, makes the task of recognising entities particularly difficult for full-text off-the-shelf Named Entity Recognition (NER) (Eiselt & Figueroa, 2013). We have utilised NER by selecting the 20 most frequent proper nouns from each of the FDB company sub-forums. A proper noun being defined as "a name used for an individual person, place, or organisation, spelt with an initial capital letter". This allows us to capture names of people and organisations being mentioned in user posts which can then be used later to record the number of LSE-listed company FDB proper nouns present in the tweets.

6.1.3.2. Stop word Removal. The removal of stop words in the tweets, FDB posts, and Reuters company descriptions has been performed using Python's NLTK package,[5] which includes a pre-built corpus of common English stop words which we use to perform stop word removal from each data source.

6.1.3.3. Lemmatisation. The NLTK has also been utilised to perform lemmatisation on the Reuters company descriptions and all of the tweets' text in order to reduce the number of words, allowing us to reduce the sparsity of our bag of words (discussed in Section 8.2.1) (Jivani, 2016).

6.2. Data fusion challenges

One of the key challenges present in this data fusion process is the heterogeneity of the three data sources. Reuters descriptions are static in the nature that this description will likely stay the same for years. FDB posts are dynamic in the sense that investors will likely be discussing recent news and events relating to a specific company.

As our approach relies on freely-available public data sources, there is the added risk that any of these data sources could suddenly become unavailable, meaning alternative features from other sources may need to be relied upon. Web scraping techniques in particular are susceptible to failing should the structure of a web page change. Utilising services which provide structured data, such as AlphaVantage, also run the risk of service shortages or their associated APIs becoming unavailable or deprecated.

Each of the data sources considered for this experiment do have reliable alternatives. Descriptions for companies can also be obtained from other reputable financial market news providers, such as Bloomberg. There are also other FDBs which do focus specifically on the LSE, although the structure for scraping posts from this FDB is significantly more challenging due to the way the websites structures its web pages. Share prices from AlphaVantage could also be obtained from web scraping, although share prices obtained in this way would likely be outdated when compared to real-time market prices.

In the next section, we perform a high-level exploratory data analysis of the collected data in order to better understand the nuances of the dataset of tweets and FDB posts.

7. Exploratory data analysis

This section will present a high-level overview of the Twitter and London South East datasets. This analysis is based on all of the tweets and FDB posts gathered for the experiment companies (Appendix A). The goal of this exploratory data analysis is to gain a better understanding of the scale of cashtag collisions, in addition to identifying any particular nuances present in the dataset which may be of importance in the annotation process (Section 8.1).

7.1. Twitter

We begin by exploring the Twitter dataset with an exploration of the cashtags within the tweets. A total of 86,539 Tweets have been collected over a one-month period from 16th April 2018 to 16th May 2018.

Taking into account the full twitter dataset of 86,539 tweets, we begin the analysis by checking how many tweets contain a cashtag which collide with one of the exchanges in Table 3. In total, 55,543 (64.2%) contain a colliding cashtag (based on our definition

[5] https://www.nltk.org/.

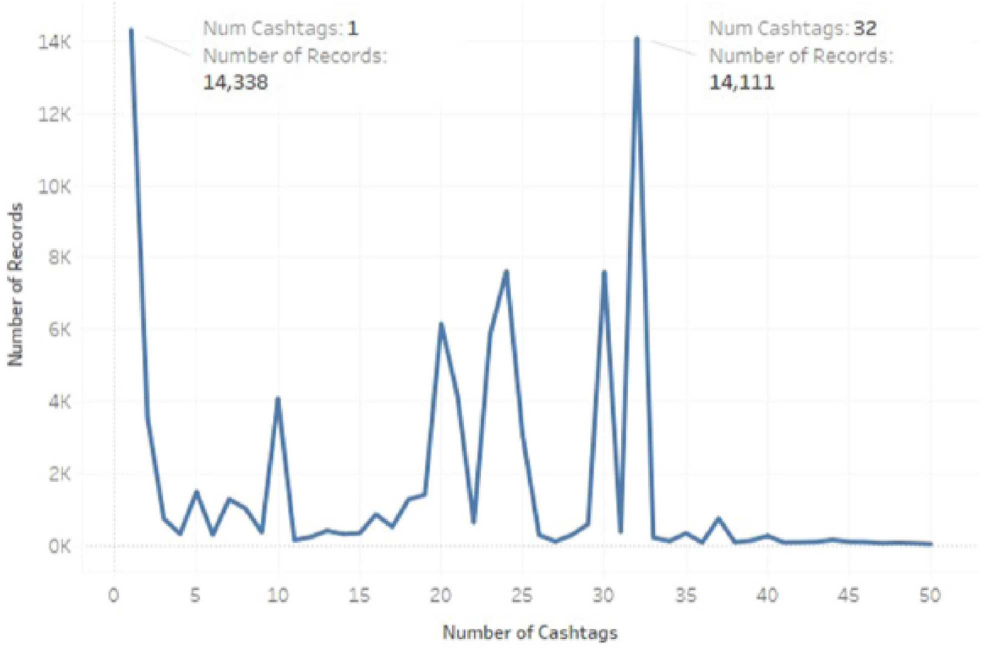


Fig. 3. Cashtag distribution.

in Section 1). This highlights the scale of the problem, which this research is attempting to address.

7.1.1. Cashtag distribution

The number of cashtags present within the tweets in our dataset falls between 1 and 50 (Fig. 3), with significant hikes at 10, 20, 24, 30, and a dramatic increase at 32 which almost exceeds that of tweets containing a single cashtag.

It is a reasonable assumption that the majority of tweets should contain one cashtag, as tweets are limited to 280 characters, allowing only a limited amount of information to be shared. There is no immediate indication as to why there is such a surge of tweets containing 32 cashtags.

7.1.2. Irregular cashtag – BTG

The most dominant cashtag in our dataset is $BTG (Fig. 4), present 58,733 times (tweets can contain duplicate cashtags). A large portion of these BTG tweets (13,309) contain the exact same textual content when not considering hyperlinks embedded within them (Fig. 5), indicating the presence of tweets created by bots. All of these tweets contain 32 cashtags, which explains the hike of cashtag distribution in Fig. 3.

The most frequent word found in BTG tweets (“binance”) refers to Binance Coin, a cryptocurrency which is currently ranked in the top twenty of all cryptocurrencies in terms of market capitalisation. There are currently over 1600 cryptocurrencies according to CoinMarketCap,[6] all of which feature their own symbol which can be converted into a cashtag on Twitter, similar to stock market ticker symbols.

The Twitter streaming API provides a structured JSON object for each tweet which contains details relating to the tweet, author, location, amongst other items. A useful attribute for detecting how a tweet was published to Twitter is the *source* field, which provides the medium used to publish a tweet.

A breakdown the most popular Tweet sources in our dataset (Fig. 6) shows a clear presence of unofficial apps generating tweets. We can now therefore conclude that the popularity of BTG cashtag in our dataset is due to the prevalence of automated cryptocurrency bots on Twitter, and that other cashtags may also be susceptible to such noise.

As a substantial number of tweets come from automated bots, this leads to a considerable amount of noise in our dataset. We do not remove these tweets from our dataset, as these tweets are clearly not related to any specific exchange, meaning the word patterns used can be of use when attempting to classify a tweet as being related to a specific exchange or not.

7.2. Financial Discussion Board (London South East) posts

Analysis of London South East company forums is significantly easier to undertake when compared to tweets, as each sub-forum is dedicated to a particular company listed on the LSE, meaning investors choose a sub-forum to discuss a specific company, thus collisions cannot exist in this domain.

7.2.1. Sector posts

The average number of posts per user of the experiment companies (Fig. 7) shows that companies listed on the AIM feature more active discussions across most sectors than their MM counterparts.

Armed with a better understanding of the Twitter and London South East datasets, the next section will introduce the methodology of resolving cashtag collisions.

8. Cashtag collision resolution methodology

The methodology of determining if a tweet contains a colliding cashtag (Fig. 8) involves the vectorisation of the tweet text into a sparse vector (Feature 1 – F1) and combining other supplementary features such as the number of exchange-specific (F2) & non-exchange-specific cashtags (F3), the count of Reuters company description words (F4), and FDB words (F5) found within the tweet so that traditional machine learning classifiers can make correlations between these features. We now proceed with the different

[6] https://coinmarketcap.com/.

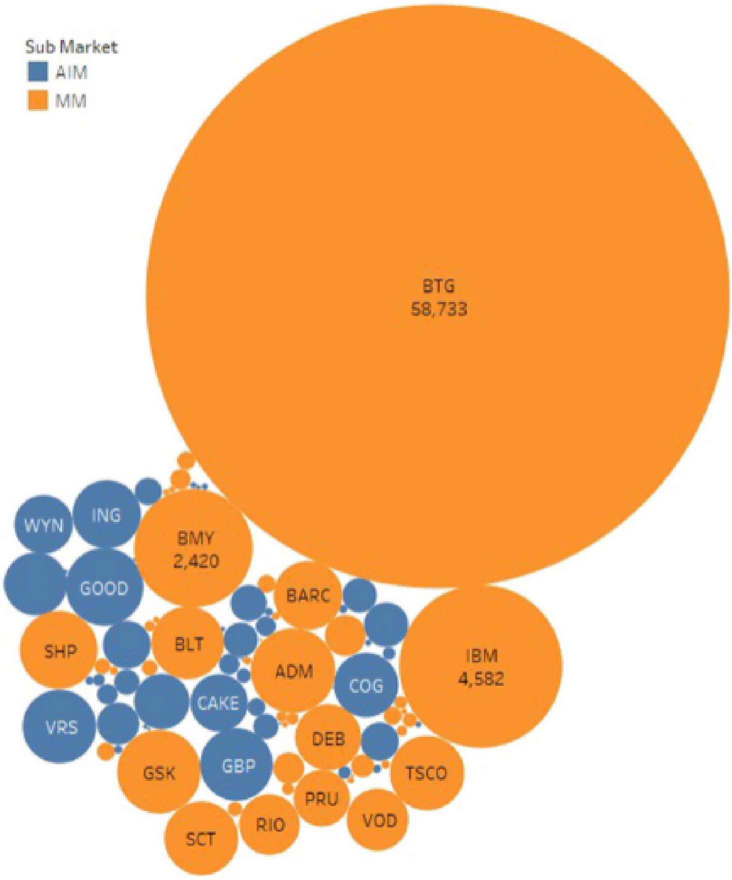
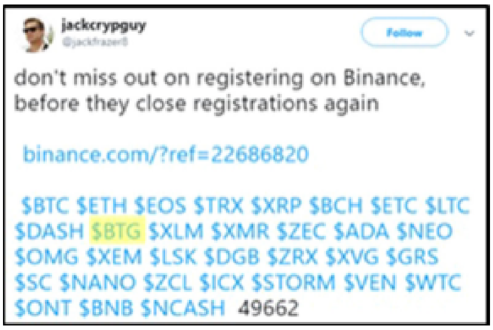


Fig. 4. BTG cashtag dominance.



don't miss out on registering on Binance, before they close registrations again

binance.com/?ref=22686820

$BTC $ETH $EOS $TRX $XRP $BCH $ETC $LTC $DASH $BTG $XLM $XMR $ZEC $ADA $NEO $OMG $XEM $LSK $DGB $ZRX $XVG $GRS $SC $NANO $ZCL $ICX $STORM $VEN $WTC $ONT $BNB $NCASH 49662

Fig. 5. Suspected bot tweet.

steps in which we detect and resolve a cashtag collision, beginning with an explanation of our annotated tweet dataset.

8.1. Annotated tweet dataset

In order to answer RQ1&2 (Section 2.3), a labelled dataset of tweets must be created in order to assess the predictive power of the different machine learning classifiers to be trained in Section 9.3. As the cost of creating a manually labelled dataset is time-consuming, particularly when the labelling requires the inspection of each tweet's text and author details, we have manually annotated 1000 tweets with the labels listed in Table 7. Although this is a laborious task even for a relatively small corpus of tweets, this is consistent with previous works relating to tweet annotation (Matsuda, Sasaki, Okazaki, & Inui, 2017; Tjong Kim Sang & van den Bosch, 2013). As the exploratory data analysis showed a heavy presence of cryptocurrency-related tweets, we use three labels to annotate our dataset. A label of zero (0) indicates the tweet does relate to a stock exchange, but not directly to the LSE. A label of one (1) indicates that the tweet directly relates to a company listed on the LSE. A label of two (2) indicates that the tweet references cryptocurrency. In order to ensure consistency in this annotation process, and to ensure high-quality labels (Abraham et al., 2016) are generated, all of these tweets have been manually annotated by a single individual experienced with annotating tweets.

8.1.1. Tweet selection

As evident from the exploratory analysis of the tweets in Section 6, the sheer dominance of the BTG cashtag means that any random selection of tweets will favour tweets containing the BTG cashtag, meaning the classifiers would generalise towards cryptocurrency tweets. To ensure fairness when selecting the 1000 tweets, we first attempt to collect ten tweets for every experiment company ticker (Appendix A). This provided 767 tweets (as some company tickers are not as actively used in tweets compared to others), for the remainder, we collect a random sample of tweets over the one-month time period for a total of 1000 tweets.

8.2. Steps 1–3: Feature design choices

We now provide a motivation for the features used to train the classifiers. Beginning with the sparse vector to represent the text of each tweet.

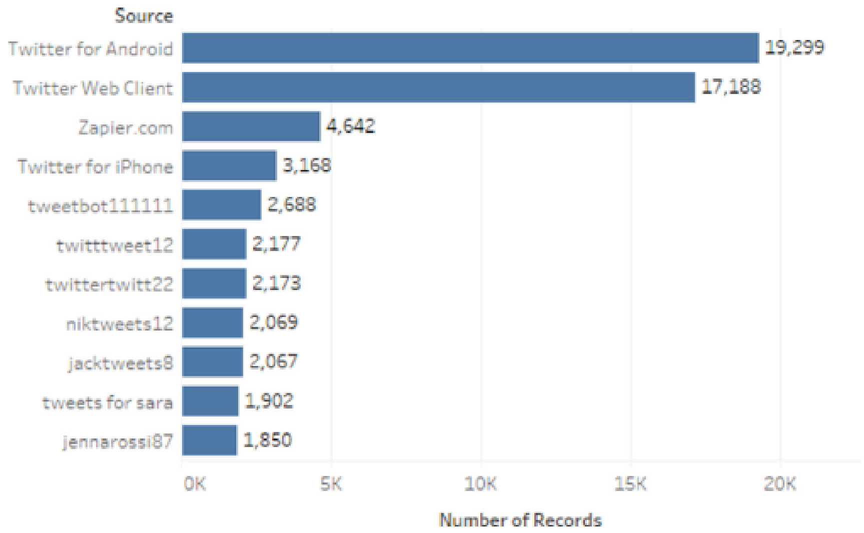


Fig. 6. Tweet sources.

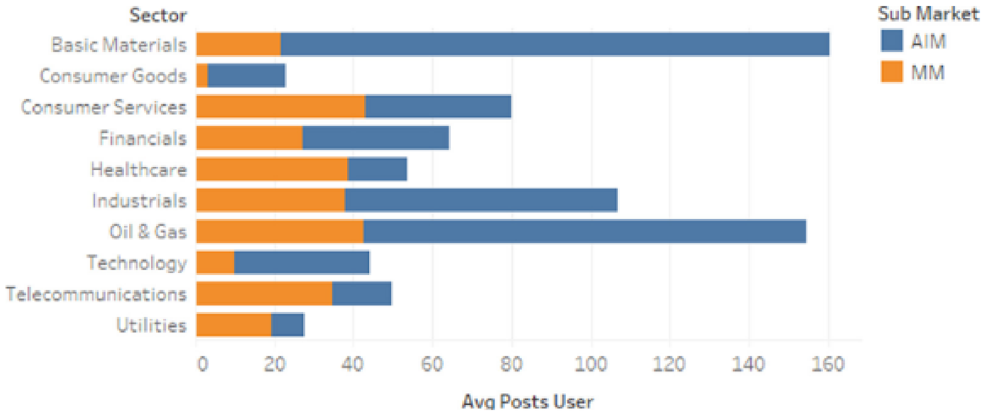


Fig. 7. Average number of posts per user (by sector).

Table 7
Annotated tweet examples.

Label	Tweet Type	Example Tweet
0	Non-LSE related	Cabot Oil & Gas Co. $COG Forecasted to Earn Q1 2018 Earnings of $0.32 Per Share
1	LSE related	Game Digital PLC 55.7% Potential Upside Indicated by Liberum Capital - - $GMD
2	Cryptocurrency related	Sign Up And Recieve 5 (LEGIT) Legitcoin tokens ($10) will be $350 $BTG $ETH $LTC $NXC 2026

8.2.1. Feature 1 (F1) – Sparse vector of tweet text

The first stage of our proposed methodology involves the conversion of all of the tweet text into a sparse matrix. After the removal of stop words and performing lemmatisation, the dimension of our sparse matrix is 1000 × 1860. This sparse matrix is featured in the training of both classifiers. As the cashtags themselves are treated as words, the classifiers will be able to make correlations between the different kinds of cashtags present within a tweet.

In regard to performing such NLP tasks on tweets in preparation for the machine learning classifiers, we elected to use the more general Python NLTK to perform this task. Although Twitter NLP-trained models do exist, none of these models have been trained to deal with the nuances present in our dataset. Although the related research (Pinto et al., 2016) surrounding NLP on tweets found that the performance of standard toolkits (such as NLTK) do not perform as well as Twitter NLP-trained models, this research did not take into account tweets relating to stock discussion, where low-character words such as stock symbols and floating-point numbers are particularly prevalent.

8.2.2. Features 2 & 3 (F2 & F3) – Count of LSE & Non-LSE cashtags in tweet

The number of exchange & non-exchange cashtags present within a tweet can be a strong indication as to whether that tweet relates to a company listed on a given exchange. If a tweet contains one cashtag which relates to the LSE, but also contains a large amount of other cashtags not listed on the LSE, this will undoubtedly assist the classification of such a tweet as being non-LSE related. As all of our tweets contain at least one LSE cashtag, the count of LSE cashtags will always be a minimum of one. As is evident from the exploratory analysis in the preceding section, cryptocurrency tweets have a substantially higher count of cashtags in them.

We have downloaded a list of all ticker symbols relating to the experiment companies listed in Table 3. We then cross-check each

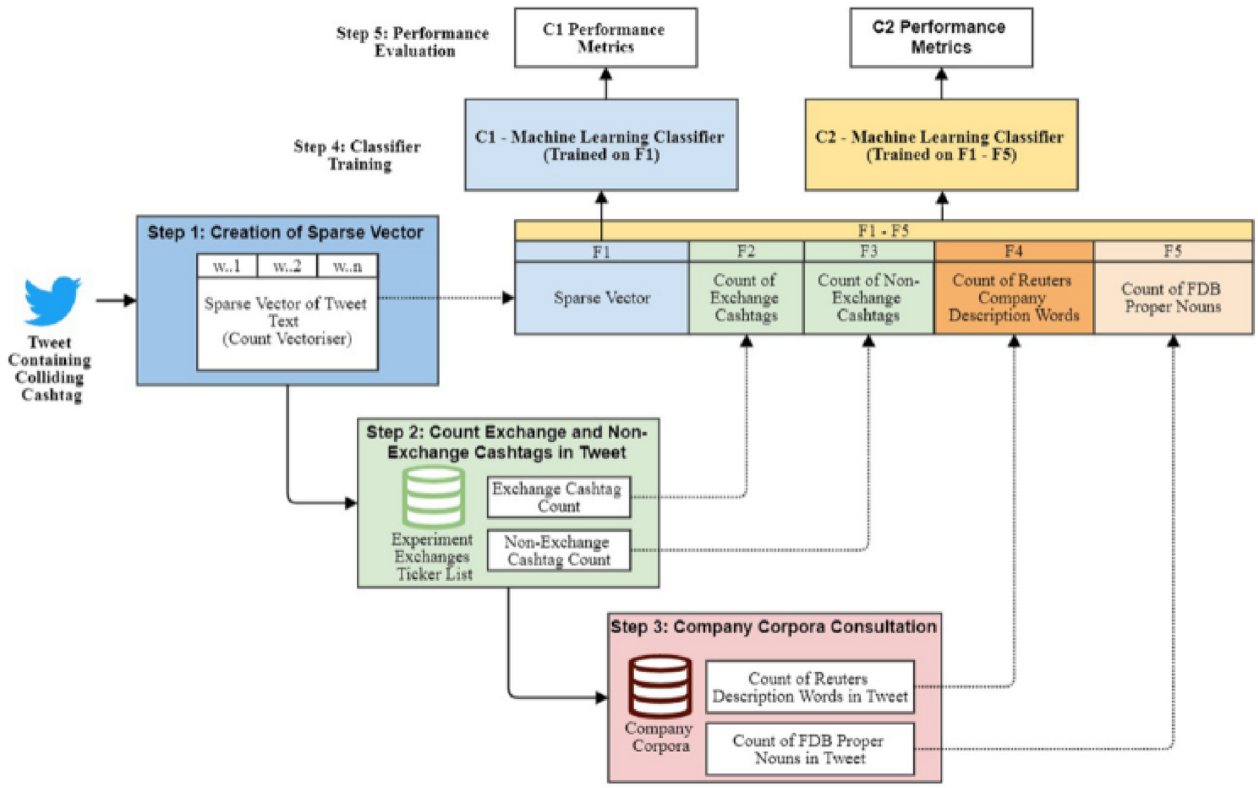
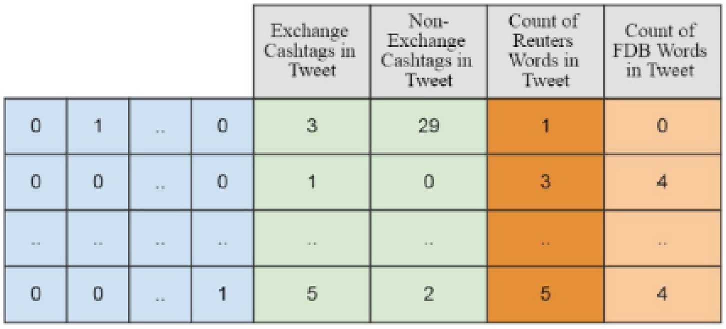

Step 5: Performance Evaluation
C1 Performance Metrics
C2 Performance Metrics
Step 4: Classifier Training
C1 - Machine Learning Classifier (Trained on F1)
C2 - Machine Learning Classifier (Trained on F1 - F5)
Step 1: Creation of Sparse Vector
w..1
w..2
w..n
Sparse Vector of Tweet Text (Count Vectoriser)
Tweet Containing Colliding Cashtag
F1 - F5
F1
F2
F3
F4
F5
Sparse Vector
Count of Exchange Cashtags
Count of Non-Exchange Cashtags
Count of Reuters Company Description Words
Count of FDB Proper Nouns
Step 2: Count Exchange and Non-Exchange Cashtags in Tweet
Experiment Exchanges Ticker List
Exchange Cashtag Count
Non-Exchange Cashtag Count
Step 3: Company Corpora Consultation
Company Corpora
Count of Reuters Description Words in Tweet
Count of FDB Proper Nouns in Tweet

Fig. 8. Cashtag collision resolution methodology.



				Exchange Cashtags in Tweet	Non-Exchange Cashtags in Tweet	Count of Reuters Words in Tweet	Count of FDB Words in Tweet
0	1	..	0	3	29	1	0
0	0	..	0	1	0	3	4
..	..	..	..	..	..	..	..
0	0	..	1	5	2	5	4

Fig. 9. Final sparse matrix representation.

tweet to see how many cashtags within the tweet relate to an LSE-listed company, with the remainder of cashtags being non-LSE cashtags.

8.2.3. Feature 4 (F4) – Count of Reuters description keywords in tweet

The count of words in the tweets which also feature in the tweet's corresponding company corpus can provide strong evidence that a tweet relates to the LSE-listed company. As low-value words have been removed from the description prior to being stored within a company's corpus, words found within the tweet text which also feature in the company description can provide a high correlation that the LSE-listed company is being referenced in the tweet. The LON:TSCO corpus, for example, features words which are able to distinguish it from its colliding company on the NASDAQ, such as "food", "retail", and "united kingdom", which would not be commonly found in tweets referencing the Tractor Supply Company.

Naturally, if two or more companies with a colliding cashtag belong to a similar sector, then this feature of counting the number of word occurrences will not provide as much value. For example, LSE:ABC (Abcam PLC) and NYSE:ABC (AmerisourceBergen Corporation) are both in the Healthcare sector, meaning their respective Reuters biographies will contain similar terminology. To alleviate this, a feature which relies on user-generated terms could be of use, this is our motivation for our final feature.

8.2.4. Feature (F5) – Count of FDB proper nouns in tweet

The final feature we have proposed is to use the most frequent proper nouns found within the FDB posts for each of the LSE-listed companies. The number of FDB proper nouns contained within the tweets could be a helpful indication to establish if a tweet refers to a specific exchange-listed company or not. The sub-forum for Tesco (LSE), for example, has frequently-discussed proper nouns such as Lidl and Aldi – Tesco's chief competitors, allowing a further distinction between LON:TSCO and NASDAQ:TSCO. This feature will be particularly more helpful to solve the more complex collisions in which two or more companies with the same ticker have the same company name but are listed on different exchanges.

In respect to these five features, we believe that, when combined (Fig. 9), they provide a more robust approach to detect a colliding cashtag tweet, versus using any single feature in isolation.

8.3. Step 4: Classifier training

After a tweet has been represented numerically by transforming it into a sparse vector, and the count of LSE, Non-LSE, Reuters, and FDB keywords have been recorded, this can then be used to train the classifiers. Based on previous works which have seen varying levels of success (Verma Scholar, Professor, & Sofat, 2014), we have chosen to train Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Naïve Bayes, Decision Tree, and Random Forest classifiers. These are each discussed in Section 9. Each of the aforementioned classifiers is trained and tested twice independently. The first classifier (C1) is trained on just the sparse vector of the tweet text (F1) alone, and the second classifier (C2) is trained on the sparse vector and other supplementary features (F1–F5) contained within the company corpora.

8.4. Step 5: Performance evaluation

The final stage of our proposed methodology involves comparing each of the classifiers to determine if a classifier benefits from being trained on the additional features. We compare the performance between the classifiers using the Matthews Correlation Coefficient score, a metric used to assess the performance of a binary classifier which has a class imbalance, discussed in further detail in Section 9.2.

The next section contains the results and discussion of the experiment results.

9. Results and discussion

This section will explore if the consideration of additional features improves the classification performance over the traditional approach of using a sparse vector alone.

The classification of tweets in this experiment is a binary classification problem – a tweet either relates to the LSE (1), or it does not (0). All of the cryptocurrency tweets (labelled 2) have been labelled zero for the training of all of the classifiers. This section will introduce a number of suitable supervised machine learning classifiers, along with their respective benefits, drawbacks, and performance on the annotated dataset.

9.1. Accuracy paradox

Before delving into each of the classifiers used in this experiment, it is important to note why we do not blindly depend on the accuracy of the models as an indication of their respective performance. High accuracy scores can often be misleading as to the predictive power of a classifier. A binary classification problem which features a dominant label can often lead to a misleading accuracy score. In our labelled dataset of 1000 tweets, 642 tweets do not correspond to the LSE, hence being labelled zero. This means if we choose to abandon our machine learning models and predict zero every time, we would achieve a 64% accuracy for free, giving a false indication of predictive power, referred to as the accuracy paradox (Valverde-albacete & Pela, 2014).

9.2. Matthews Correlation Coefficient

A more practical approach to evaluating the results of a binary classifier in which there is class imbalance is the Matthews Correlation Coefficient (MCC) (Boughorbel, Jarray, & El-Anbari, 2017).

Table 8
Logistic Regression results.

	Sparse Vector		Combined Features	
CM	616	26	618	24
	50	308	40	318
MCC Score	0.83		0.86	

Table 9
kNN results.

	Sparse Vector		Combined Features	
CM	609	33	588	54
	73	285	58	300
MCC Score	0.77		0.76	

The MCC score (Eq. (1)) is calculated by using the Confusion Matrix (CM) results using the equation below (where TP = true positive, TN = true negative, FP = false positive, and FN = false negative):

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (1)$$

The MCC score returns a value from −1 to +1. A value of +1 indicates the model makes perfect predictions, 0 indicates the model is no better than random chance, with −1 representing the classifier has made incorrect predictions across the board (Liu, Cheng, Yan, Wu, & Chen, 2015).

Once each of the classifiers' performance has been discussed, we compare the two best performing classifiers (in respect to their MCC score), to determine if the results between the two best performers are statistically significant. H_0 denotes the null hypothesis, which we will attempt to reject at a significance level of five percent. H_1 denotes the alternative hypothesis, which we will attempt to lend support to if we are able to reject H_0.

$$H_0 : MCC_{c_1} < MCC_{c_2}$$
$$H_1 : MCC_{c_1} \geq MCC_{c_2}$$

9.3. Machine learning classifiers

All of the classifiers have been implemented using the skikit-learn library within Python. Each classification model has differing hyperparameters which can affect the performance metrics of the classifier, we find optimal hyperparameters for each classifier through the use of a grid search, which explores a user-specified parameter space to determine the most efficient combination of hyperparameters in respect to a scoring metric (we elect to choose the best hyperparameter combinations based on the MCC score) (Öğüt, Mete Doğanay, & Aktaş, 2009). A common approach suggested by Geron (2017) is to start with a coarse grid search covering a wide parameter space, and then a finer grid search based on the best values found – we have adopted this approach. Internal 10k-fold cross validation has been used for each classifier using an 80/20 train/test split.

A complete table of results for each classifier is provided in Table 14.

9.3.1. Logistic Regression

The first classifier we consider is Logistic Regression (LR), due to its suitability for relatively small training sets (Perlich, Provost, Simonoff, & Stern, 2003). The LR results (Table 8) show an observable increase in the MCC score when the classifier is trained on the combined features when compared to just the sparse vector alone.

9.3.2. K-Nearest Neighbours

The next classifier trained is the K-Nearest Neighbours (kNN) classifier. The kNN results (Table 9) show that the classifier trained on the combined features does not yield a better MCC score compared to the sparse vector alone.

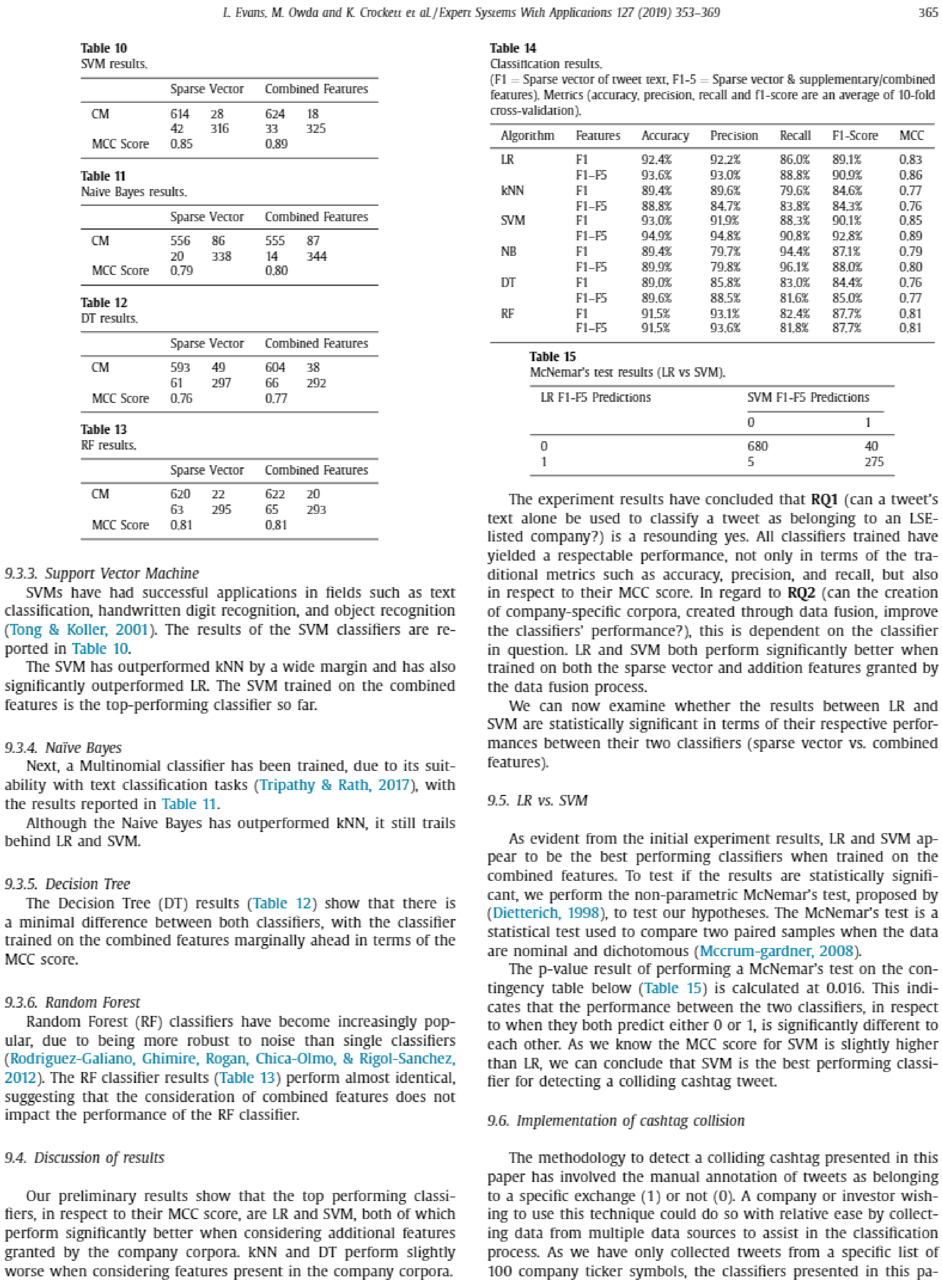


Table 10
SVM results.

	Sparse Vector		Combined Features	
CM	614	28	624	18
	42	316	33	325
MCC Score	0.85		0.89	

Table 11
Naive Bayes results.

	Sparse Vector		Combined Features	
CM	556	86	555	87
	20	338	14	344
MCC Score	0.79		0.80	

Table 12
DT results.

	Sparse Vector		Combined Features	
CM	593	49	604	38
	61	297	66	292
MCC Score	0.76		0.77	

Table 13
RF results.

	Sparse Vector		Combined Features	
CM	620	22	622	20
	63	295	65	293
MCC Score	0.81		0.81	

9.3.3. *Support Vector Machine*

SVMs have had successful applications in fields such as text classification, handwritten digit recognition, and object recognition (Tong & Koller, 2001). The results of the SVM classifiers are reported in Table 10.

The SVM has outperformed kNN by a wide margin and has also significantly outperformed LR. The SVM trained on the combined features is the top-performing classifier so far.

9.3.4. *Naïve Bayes*

Next, a Multinomial classifier has been trained, due to its suitability with text classification tasks (Tripathy & Rath, 2017), with the results reported in Table 11.

Although the Naive Bayes has outperformed kNN, it still trails behind LR and SVM.

9.3.5. *Decision Tree*

The Decision Tree (DT) results (Table 12) show that there is a minimal difference between both classifiers, with the classifier trained on the combined features marginally ahead in terms of the MCC score.

9.3.6. *Random Forest*

Random Forest (RF) classifiers have become increasingly popular, due to being more robust to noise than single classifiers (Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012). The RF classifier results (Table 13) perform almost identical, suggesting that the consideration of combined features does not impact the performance of the RF classifier.

9.4. *Discussion of results*

Our preliminary results show that the top performing classifiers, in respect to their MCC score, are LR and SVM, both of which perform significantly better when considering additional features granted by the company corpora. kNN and DT perform slightly worse when considering features present in the company corpora.

Table 14
Classification results.
(F1 = Sparse vector of tweet text, F1-5 = Sparse vector & supplementary/combined features). Metrics (accuracy, precision, recall and f1-score are an average of 10-fold cross-validation).

Algorithm	Features	Accuracy	Precision	Recall	F1-Score	MCC
LR	F1	92.4%	92.2%	86.0%	89.1%	0.83
	F1–F5	93.6%	93.0%	88.8%	90.9%	0.86
kNN	F1	89.4%	89.6%	79.6%	84.6%	0.77
	F1–F5	88.8%	84.7%	83.8%	84.3%	0.76
SVM	F1	93.0%	91.9%	88.3%	90.1%	0.85
	F1–F5	94.9%	94.8%	90.8%	92.8%	0.89
NB	F1	89.4%	79.7%	94.4%	87.1%	0.79
	F1–F5	89.9%	79.8%	96.1%	88.0%	0.80
DT	F1	89.0%	85.8%	83.0%	84.4%	0.76
	F1–F5	89.6%	88.5%	81.6%	85.0%	0.77
RF	F1	91.5%	93.1%	82.4%	87.7%	0.81
	F1–F5	91.5%	93.6%	81.8%	87.7%	0.81

Table 15
McNemar's test results (LR vs SVM).

LR F1-F5 Predictions	SVM F1-F5 Predictions	
	0	1
0	680	40
1	5	275

The experiment results have concluded that **RQ1** (can a tweet's text alone be used to classify a tweet as belonging to an LSE-listed company?) is a resounding yes. All classifiers trained have yielded a respectable performance, not only in terms of the traditional metrics such as accuracy, precision, and recall, but also in respect to their MCC score. In regard to **RQ2** (can the creation of company-specific corpora, created through data fusion, improve the classifiers' performance?), this is dependent on the classifier in question. LR and SVM both perform significantly better when trained on both the sparse vector and addition features granted by the data fusion process.

We can now examine whether the results between LR and SVM are statistically significant in terms of their respective performances between their two classifiers (sparse vector vs. combined features).

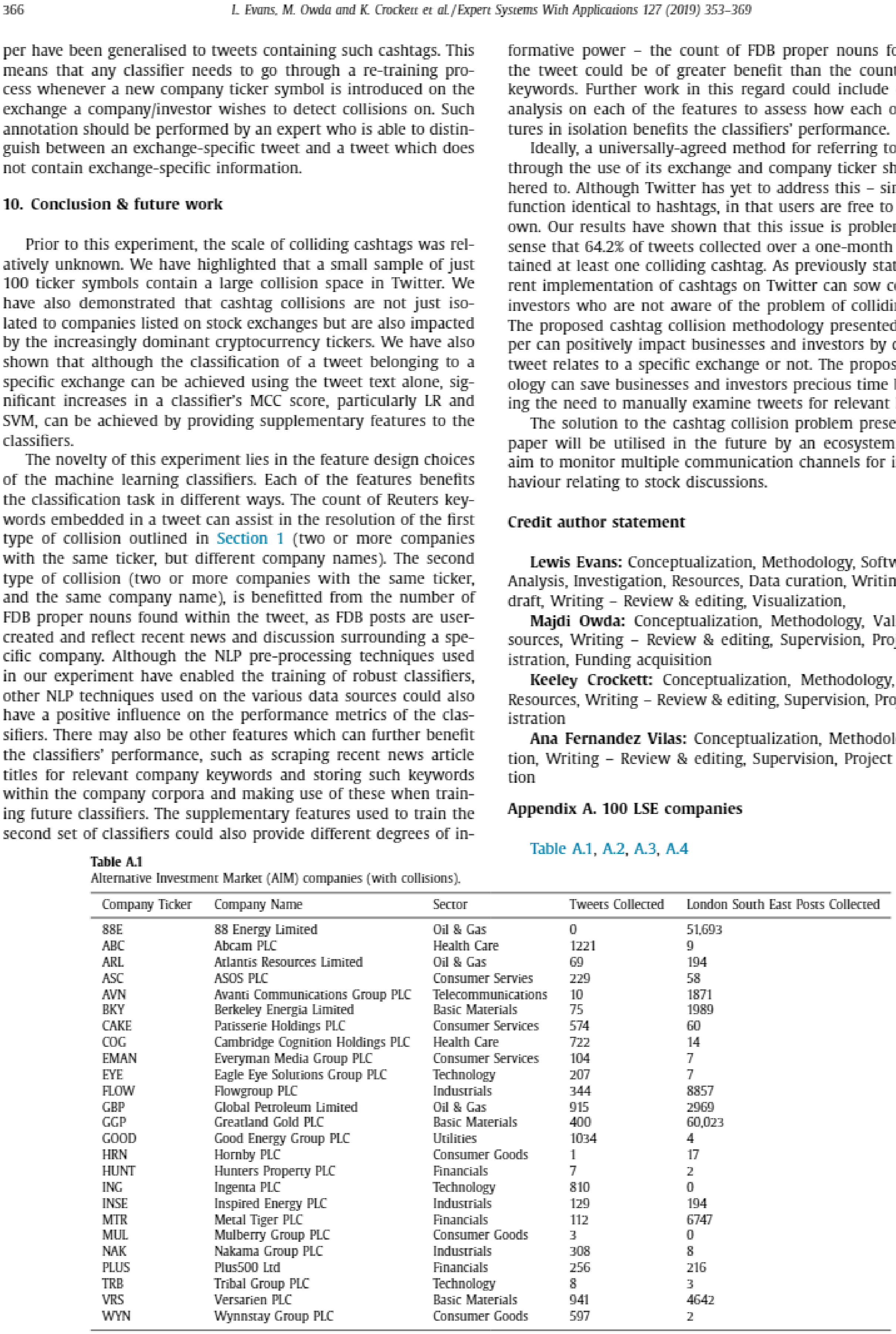
9.5. *LR vs. SVM*

As evident from the initial experiment results, LR and SVM appear to be the best performing classifiers when trained on the combined features. To test if the results are statistically significant, we perform the non-parametric McNemar's test, proposed by (Dietterich, 1998), to test our hypotheses. The McNemar's test is a statistical test used to compare two paired samples when the data are nominal and dichotomous (Mccrum-gardner, 2008).

The p-value result of performing a McNemar's test on the contingency table below (Table 15) is calculated at 0.016. This indicates that the performance between the two classifiers, in respect to when they both predict either 0 or 1, is significantly different to each other. As we know the MCC score for SVM is slightly higher than LR, we can conclude that SVM is the best performing classifier for detecting a colliding cashtag tweet.

9.6. *Implementation of cashtag collision*

The methodology to detect a colliding cashtag presented in this paper has involved the manual annotation of tweets as belonging to a specific exchange (1) or not (0). A company or investor wishing to use this technique could do so with relative ease by collecting data from multiple data sources to assist in the classification process. As we have only collected tweets from a specific list of 100 company ticker symbols, the classifiers presented in this pa-



per have been generalised to tweets containing such cashtags. This means that any classifier needs to go through a re-training process whenever a new company ticker symbol is introduced on the exchange a company/investor wishes to detect collisions on. Such annotation should be performed by an expert who is able to distinguish between an exchange-specific tweet and a tweet which does not contain exchange-specific information.

10. Conclusion & future work

Prior to this experiment, the scale of colliding cashtags was relatively unknown. We have highlighted that a small sample of just 100 ticker symbols contain a large collision space in Twitter. We have also demonstrated that cashtag collisions are not just isolated to companies listed on stock exchanges but are also impacted by the increasingly dominant cryptocurrency tickers. We have also shown that although the classification of a tweet belonging to a specific exchange can be achieved using the tweet text alone, significant increases in a classifier's MCC score, particularly LR and SVM, can be achieved by providing supplementary features to the classifiers.

The novelty of this experiment lies in the feature design choices of the machine learning classifiers. Each of the features benefits the classification task in different ways. The count of Reuters keywords embedded in a tweet can assist in the resolution of the first type of collision outlined in Section 1 (two or more companies with the same ticker, but different company names). The second type of collision (two or more companies with the same ticker, and the same company name), is benefitted from the number of FDB proper nouns found within the tweet, as FDB posts are user-created and reflect recent news and discussion surrounding a specific company. Although the NLP pre-processing techniques used in our experiment have enabled the training of robust classifiers, other NLP techniques used on the various data sources could also have a positive influence on the performance metrics of the classifiers. There may also be other features which can further benefit the classifiers' performance, such as scraping recent news article titles for relevant company keywords and storing such keywords within the company corpora and making use of these when training future classifiers. The supplementary features used to train the second set of classifiers could also provide different degrees of informative power – the count of FDB proper nouns found within the tweet could be of greater benefit than the count of Reuters keywords. Further work in this regard could include quantitative analysis on each of the features to assess how each of these features in isolation benefits the classifiers' performance.

Ideally, a universally-agreed method for referring to a company through the use of its exchange and company ticker should be adhered to. Although Twitter has yet to address this – since cashtags function identical to hashtags, in that users are free to create their own. Our results have shown that this issue is problematic in the sense that 64.2% of tweets collected over a one-month period contained at least one colliding cashtag. As previously stated, the current implementation of cashtags on Twitter can sow confusion for investors who are not aware of the problem of colliding cashtags. The proposed cashtag collision methodology presented in this paper can positively impact businesses and investors by deciding if a tweet relates to a specific exchange or not. The proposed methodology can save businesses and investors precious time by eliminating the need to manually examine tweets for relevant keywords.

The solution to the cashtag collision problem presented in this paper will be utilised in the future by an ecosystem which will aim to monitor multiple communication channels for irregular behaviour relating to stock discussions.

Credit author statement

Lewis Evans: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Resources, Data curation, Writing – Original draft, Writing – Review & editing, Visualization,

Majdi Owda: Conceptualization, Methodology, Validation, Resources, Writing – Review & editing, Supervision, Project Administration, Funding acquisition

Keeley Crockett: Conceptualization, Methodology, Validation, Resources, Writing – Review & editing, Supervision, Project Administration

Ana Fernandez Vilas: Conceptualization, Methodology, Validation, Writing – Review & editing, Supervision, Project administration

Appendix A. 100 LSE companies

Table A.1, A.2, A.3, A.4

Table A.1
Alternative Investment Market (AIM) companies (with collisions).

Company Ticker	Company Name	Sector	Tweets Collected	London South East Posts Collected
88E	88 Energy Limited	Oil & Gas	0	51,693
ABC	Abcam PLC	Health Care	1221	9
ARL	Atlantis Resources Limited	Oil & Gas	69	194
ASC	ASOS PLC	Consumer Servies	229	58
AVN	Avanti Communications Group PLC	Telecommunications	10	1871
BKY	Berkeley Energia Limited	Basic Materials	75	1989
CAKE	Patisserie Holdings PLC	Consumer Services	574	60
COG	Cambridge Cognition Holdings PLC	Health Care	722	14
EMAN	Everyman Media Group PLC	Consumer Services	104	7
EYE	Eagle Eye Solutions Group PLC	Technology	207	7
FLOW	Flowgroup PLC	Industrials	344	8857
GBP	Global Petroleum Limited	Oil & Gas	915	2969
GGP	Greatland Gold PLC	Basic Materials	400	60,023
GOOD	Good Energy Group PLC	Utilities	1034	4
HRN	Hornby PLC	Consumer Goods	1	17
HUNT	Hunters Property PLC	Financials	7	2
ING	Ingenta PLC	Technology	810	0
INSE	Inspired Energy PLC	Industrials	129	194
MTR	Metal Tiger PLC	Financials	112	6747
MUL	Mulberry Group PLC	Consumer Goods	3	0
NAK	Nakama Group PLC	Industrials	308	8
PLUS	Plus500 Ltd	Financials	256	216
TRB	Tribal Group PLC	Technology	8	3
VRS	Versarien PLC	Basic Materials	941	4642
WYN	Wynnstay Group PLC	Consumer Goods	597	2

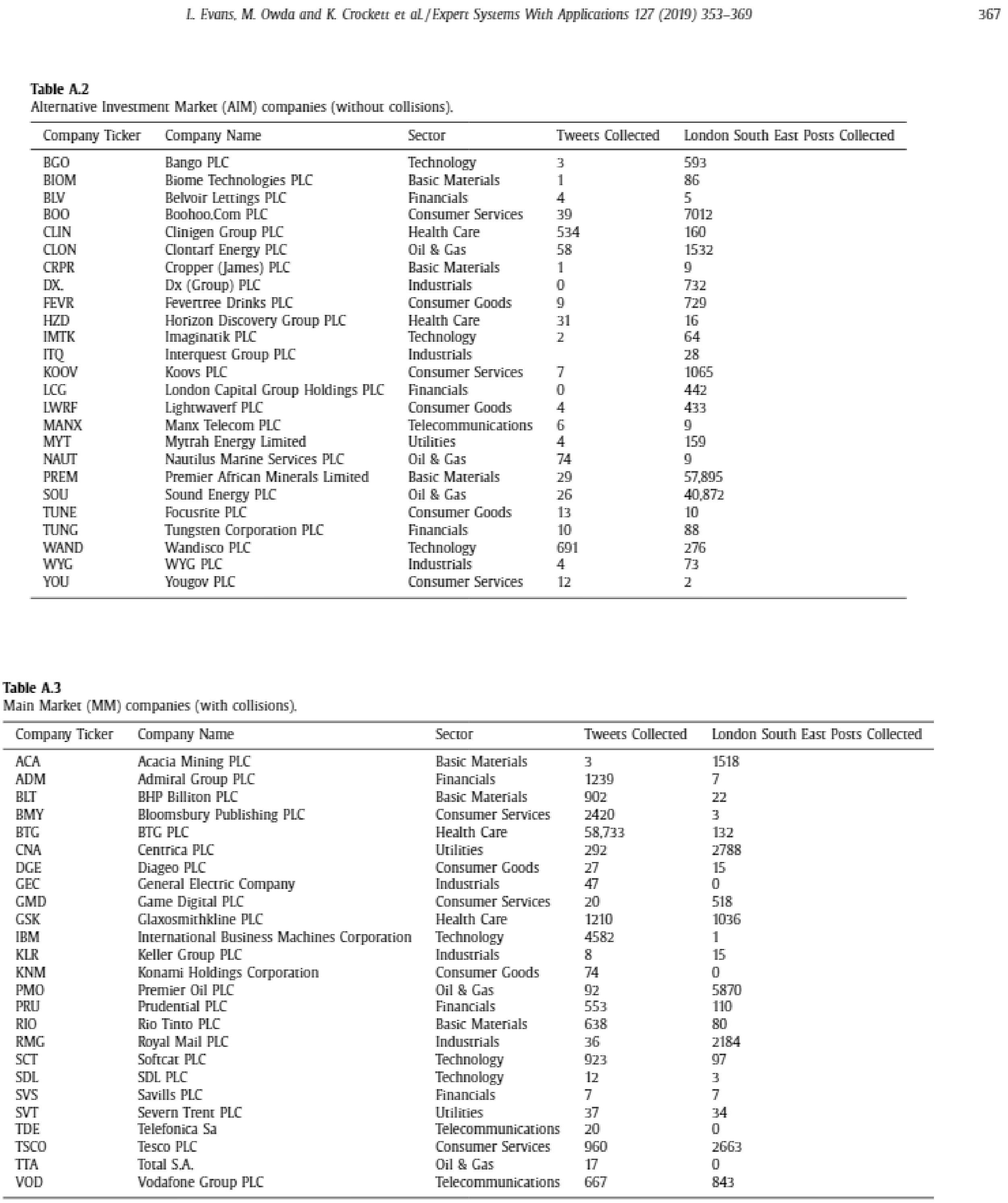


Table A.2
Alternative Investment Market (AIM) companies (without collisions).

Company Ticker	Company Name	Sector	Tweets Collected	London South East Posts Collected
BGO	Bango PLC	Technology	3	593
BIOM	Biome Technologies PLC	Basic Materials	1	86
BLV	Belvoir Lettings PLC	Financials	4	5
BOO	Boohoo.Com PLC	Consumer Services	39	7012
CLIN	Clinigen Group PLC	Health Care	534	160
CLON	Clontarf Energy PLC	Oil & Gas	58	1532
CRPR	Cropper (James) PLC	Basic Materials	1	9
DX.	Dx (Group) PLC	Industrials	0	732
FEVR	Fevertree Drinks PLC	Consumer Goods	9	729
HZD	Horizon Discovery Group PLC	Health Care	31	16
IMTK	Imaginatik PLC	Technology	2	64
ITQ	Interquest Group PLC	Industrials		28
KOOV	Koovs PLC	Consumer Services	7	1065
LCG	London Capital Group Holdings PLC	Financials	0	442
LWRF	Lightwaverf PLC	Consumer Goods	4	433
MANX	Manx Telecom PLC	Telecommunications	6	9
MYT	Mytrah Energy Limited	Utilities	4	159
NAUT	Nautilus Marine Services PLC	Oil & Gas	74	9
PREM	Premier African Minerals Limited	Basic Materials	29	57,895
SOU	Sound Energy PLC	Oil & Gas	26	40,872
TUNE	Focusrite PLC	Consumer Goods	13	10
TUNG	Tungsten Corporation PLC	Financials	10	88
WAND	Wandisco PLC	Technology	691	276
WYG	WYG PLC	Industrials	4	73
YOU	Yougov PLC	Consumer Services	12	2

Table A.3
Main Market (MM) companies (with collisions).

Company Ticker	Company Name	Sector	Tweets Collected	London South East Posts Collected
ACA	Acacia Mining PLC	Basic Materials	3	1518
ADM	Admiral Group PLC	Financials	1239	7
BLT	BHP Billiton PLC	Basic Materials	902	22
BMY	Bloomsbury Publishing PLC	Consumer Services	2420	3
BTG	BTG PLC	Health Care	58,733	132
CNA	Centrica PLC	Utilities	292	2788
DGE	Diageo PLC	Consumer Goods	27	15
GEC	General Electric Company	Industrials	47	0
GMD	Game Digital PLC	Consumer Services	20	518
GSK	Glaxosmithkline PLC	Health Care	1210	1036
IBM	International Business Machines Corporation	Technology	4582	1
KLR	Keller Group PLC	Industrials	8	15
KNM	Konami Holdings Corporation	Consumer Goods	74	0
PMO	Premier Oil PLC	Oil & Gas	92	5870
PRU	Prudential PLC	Financials	553	110
RIO	Rio Tinto PLC	Basic Materials	638	80
RMG	Royal Mail PLC	Industrials	36	2184
SCT	Softcat PLC	Technology	923	97
SDL	SDL PLC	Technology	12	3
SVS	Savills PLC	Financials	7	7
SVT	Severn Trent PLC	Utilities	37	34
TDE	Telefonica Sa	Telecommunications	20	0
TSCO	Tesco PLC	Consumer Services	960	2663
TTA	Total S.A.	Oil & Gas	17	0
VOD	Vodafone Group PLC	Telecommunications	667	843

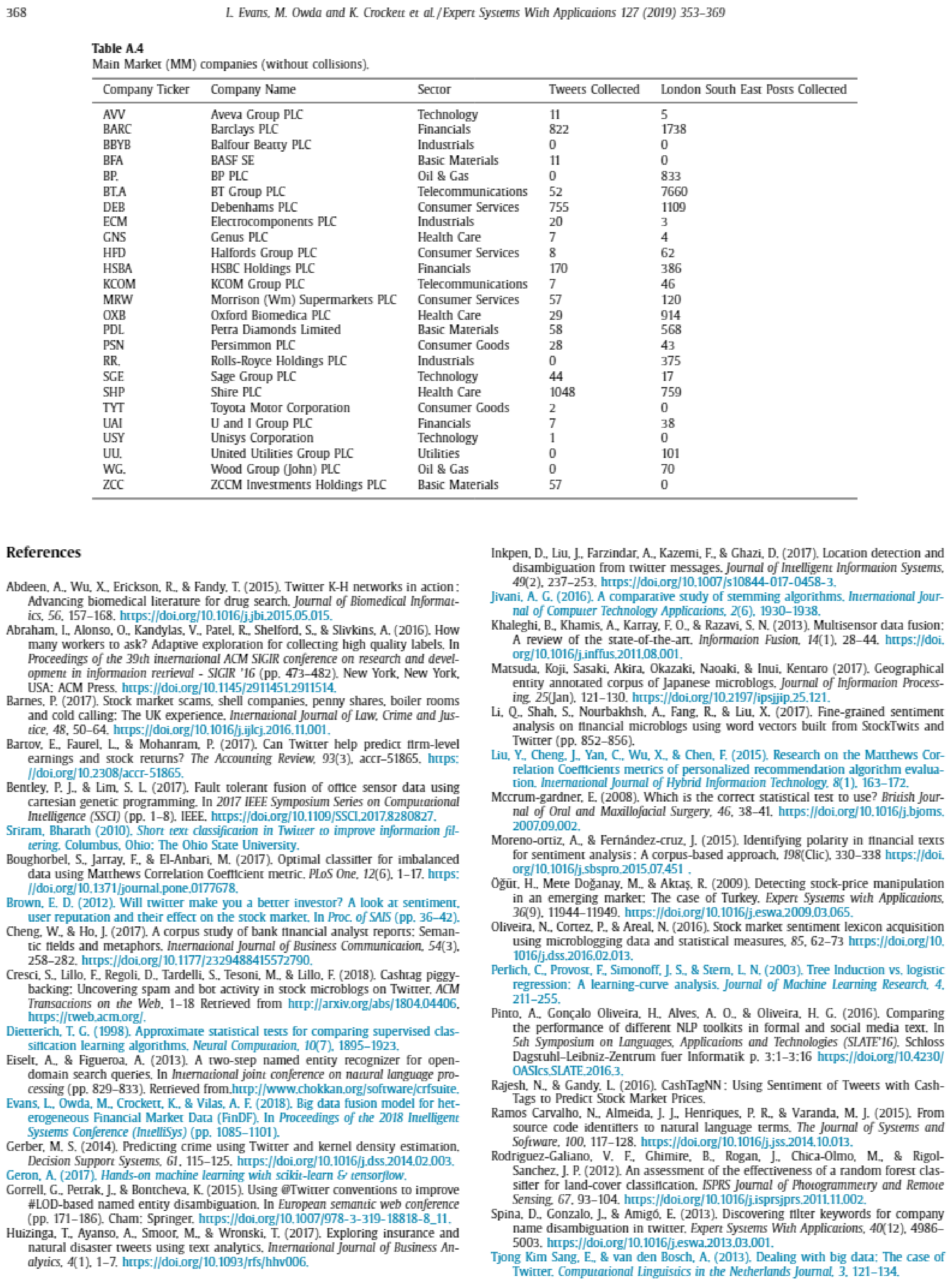


Table A.4
Main Market (MM) companies (without collisions).

Company Ticker	Company Name	Sector	Tweets Collected	London South East Posts Collected
AVV	Aveva Group PLC	Technology	11	5
BARC	Barclays PLC	Financials	822	1738
BBYB	Balfour Beatty PLC	Industrials	0	0
BFA	BASF SE	Basic Materials	11	0
BP.	BP PLC	Oil & Gas	0	833
BT.A	BT Group PLC	Telecommunications	52	7660
DEB	Debenhams PLC	Consumer Services	755	1109
ECM	Electrocomponents PLC	Industrials	20	3
GNS	Genus PLC	Health Care	7	4
HFD	Halfords Group PLC	Consumer Services	8	62
HSBA	HSBC Holdings PLC	Financials	170	386
KCOM	KCOM Group PLC	Telecommunications	7	46
MRW	Morrison (Wm) Supermarkets PLC	Consumer Services	57	120
OXB	Oxford Biomedica PLC	Health Care	29	914
PDL	Petra Diamonds Limited	Basic Materials	58	568
PSN	Persimmon PLC	Consumer Goods	28	43
RR.	Rolls-Royce Holdings PLC	Industrials	0	375
SGE	Sage Group PLC	Technology	44	17
SHP	Shire PLC	Health Care	1048	759
TYT	Toyota Motor Corporation	Consumer Goods	2	0
UAI	U and I Group PLC	Financials	7	38
USY	Unisys Corporation	Technology	1	0
UU.	United Utilities Group PLC	Utilities	0	101
WG.	Wood Group (John) PLC	Oil & Gas	0	70
ZCC	ZCCM Investments Holdings PLC	Basic Materials	57	0

References

Abdeen, A., Wu, X., Erickson, R., & Fandy, T. (2015). Twitter K-H networks in action : Advancing biomedical literature for drug search. *Journal of Biomedical Informatics, 56*, 157–168. https://doi.org/10.1016/j.jbi.2015.05.015.

Abraham, I., Alonso, O., Kandylas, V., Patel, R., Shelford, S., & Slivkins, A. (2016). How many workers to ask? Adaptive exploration for collecting high quality labels. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval - SIGIR '16* (pp. 473–482). New York, New York, USA: ACM Press. https://doi.org/10.1145/2911451.2911514.

Barnes, P. (2017). Stock market scams, shell companies, penny shares, boiler rooms and cold calling: The UK experience. *International Journal of Law, Crime and Justice, 48*, 50–64. https://doi.org/10.1016/j.ijlcj.2016.11.001.

Bartov, E., Faurel, L., & Mohanram, P. (2017). Can Twitter help predict firm-level earnings and stock returns? *The Accounting Review, 93*(3), accr–51865. https://doi.org/10.2308/accr-51865.

Bentley, P. J., & Lim, S. L. (2017). Fault tolerant fusion of office sensor data using cartesian genetic programming. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–8). IEEE. https://doi.org/10.1109/SSCI.2017.8280827.

Sriram, Bharath (2010). *Short text classification in Twitter to improve information filtering*. Columbus, Ohio: The Ohio State University.

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One, 12*(6), 1–17. https://doi.org/10.1371/journal.pone.0177678.

Brown, E. D. (2012). Will twitter make you a better investor? A look at sentiment, user reputation and their effect on the stock market. In *Proc. of SAIS* (pp. 36–42).

Cheng, W., & Ho, J. (2017). A corpus study of bank financial analyst reports: Semantic fields and metaphors. *International Journal of Business Communication, 54*(3), 258–282. https://doi.org/10.1177/2329488415572790.

Cresci, S., Lillo, F., Regoli, D., Tardelli, S., Tesoni, M., & Lillo, F. (2018). Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on Twitter. *ACM Transactions on the Web*, 1–18 Retrieved from http://arxiv.org/abs/1804.04406, https://tweb.acm.org/.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1923.

Eiselt, A., & Figueroa, A. (2013). A two-step named entity recognizer for open-domain search queries. In *International joint conference on natural language processing* (pp. 829–833). Retrieved from.http://www.chokkan.org/software/crfsuite.

Evans, L., Owda, M., Crockett, K., & Vilas, A. F. (2018). Big data fusion model for heterogeneous Financial Market Data (FinDF). In *Proceedings of the 2018 Intelligent Systems Conference (IntelliSys)* (pp. 1085–1101).

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems, 61*, 115–125. https://doi.org/10.1016/j.dss.2014.02.003.

Geron, A. (2017). *Hands-on machine learning with scikit-learn & tensorflow*.

Gorrell, G., Petrak, J., & Bontcheva, K. (2015). Using @Twitter conventions to improve #LOD-based named entity disambiguation. In *European semantic web conference* (pp. 171–186). Cham: Springer. https://doi.org/10.1007/978-3-319-18818-8_11.

Huizinga, T., Ayanso, A., Smoor, M., & Wronski, T. (2017). Exploring insurance and natural disaster tweets using text analytics. *International Journal of Business Analytics, 4*(1), 1–7. https://doi.org/10.1093/rfs/hhv006.

Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., & Ghazi, D. (2017). Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems, 49*(2), 237–253. https://doi.org/10.1007/s10844-017-0458-3.

Jivani, A. G. (2016). A comparative study of stemming algorithms. *International Journal of Computer Technology Applications, 2*(6), 1930–1938.

Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion, 14*(1), 28–44. https://doi.org/10.1016/j.inffus.2011.08.001.

Matsuda, Koji, Sasaki, Akira, Okazaki, Naoaki, & Inui, Kentaro (2017). Geographical entity annotated corpus of Japanese microblogs. *Journal of Information Processing, 25*(Jan), 121–130. https://doi.org/10.2197/ipsjjip.25.121.

Li, Q., Shah, S., Nourbakhsh, A., Fang, R., & Liu, X. (2017). Fine-grained sentiment analysis on financial microblogs using word vectors built from StockTwits and Twitter (pp. 852–856).

Liu, Y., Cheng, J., Yan, C., Wu, X., & Chen, F. (2015). Research on the Matthews Correlation Coefficients metrics of personalized recommendation algorithm evaluation. *International Journal of Hybrid Information Technology, 8*(1), 163–172.

Mccrum-gardner, E. (2008). Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery, 46*, 38–41. https://doi.org/10.1016/j.bjoms.2007.09.002.

Moreno-ortiz, A., & Fernández-cruz, J. (2015). Identifying polarity in financial texts for sentiment analysis : A corpus-based approach, *198*(Clic), 330–338 https://doi.org/10.1016/j.sbspro.2015.07.451 .

Öğüt, H., Mete Doğanay, M., & Aktaş, R. (2009). Detecting stock-price manipulation in an emerging market: The case of Turkey. *Expert Systems with Applications, 36*(9), 11944–11949. https://doi.org/10.1016/j.eswa.2009.03.065.

Oliveira, N., Cortez, P., & Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures, *85*, 62–73 https://doi.org/10.1016/j.dss.2016.02.013.

Perlich, C., Provost, F., Simonoff, J. S., & Stern, L. N. (2003). Tree Induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research, 4*, 211–255.

Pinto, A., Gonçalo Oliveira, H., Alves, A. O., & Oliveira, H. G. (2016). Comparing the performance of different NLP toolkits in formal and social media text. In *5th Symposium on Languages, Applications and Technologies (SLATE'16)*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik p. 3:1–3:16 https://doi.org/10.4230/OASIcs.SLATE.2016.3.

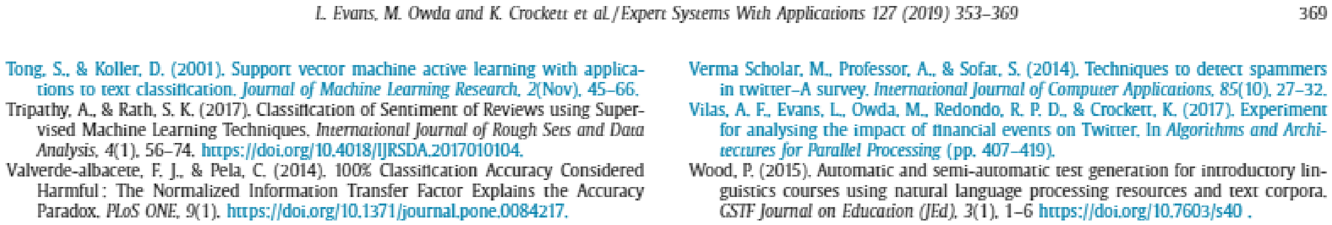
Rajesh, N., & Gandy, L. (2016). CashTagNN : Using Sentiment of Tweets with CashTags to Predict Stock Market Prices.

Ramos Carvalho, N., Almeida, J. J., Henriques, P. R., & Varanda, M. J. (2015). From source code identifiers to natural language terms. *The Journal of Systems and Software, 100*, 117–128. https://doi.org/10.1016/j.jss.2014.10.013.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing, 67*, 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002.

Spina, D., Gonzalo, J., & Amigó, E. (2013). Discovering filter keywords for company name disambiguation in twitter. *Expert Systems With Applications, 40*(12), 4986–5003. https://doi.org/10.1016/j.eswa.2013.03.001.

Tjong Kim Sang, E., & van den Bosch, A. (2013). Dealing with big data: The case of Twitter. *Computational Linguistics in the Netherlands Journal, 3*, 121–134.



Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research, 2*(Nov), 45–66.

Tripathy, A., & Rath, S. K. (2017). Classification of Sentiment of Reviews using Supervised Machine Learning Techniques. *International Journal of Rough Sets and Data Analysis, 4*(1), 56–74. https://doi.org/10.4018/IJRSDA.2017010104.

Valverde-albacete, F. J., & Pela, C. (2014). 100% Classification Accuracy Considered Harmful : The Normalized Information Transfer Factor Explains the Accuracy Paradox. *PLoS ONE, 9*(1). https://doi.org/10.1371/journal.pone.0084217.

Verma Scholar, M., Professor, A., & Sofat, S. (2014). Techniques to detect spammers in twitter–A survey. *International Journal of Computer Applications, 85*(10), 27–32.

Vilas, A. F., Evans, L., Owda, M., Redondo, R. P. D., & Crockett, K. (2017). Experiment for analysing the impact of financial events on Twitter. In *Algorithms and Architectures for Parallel Processing* (pp. 407–419).

Wood, P. (2015). Automatic and semi-automatic test generation for introductory linguistics courses using natural language processing resources and text corpora. *GSTF Journal on Education (JEd), 3*(1), 1–6 https://doi.org/10.7603/s40 .

Appendix D

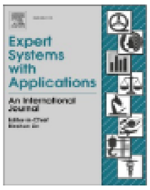
Credibility Assessment of Financial Stock Tweets Paper

Expert Systems With Applications 168 (2021) 114351

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Credibility assessment of financial stock tweets

Lewis Evans [a,*], Majdi Owda [a], Keeley Crockett [a], Ana Fernandez Vilas [b]

[a] *Department of Computing and Mathematics, Manchester Metropolitan University M1 5GD UK Manchester, England*
[b] *Ana Fernandez Vilas, I&C Lab, AtlantTIC Research Centre, University of Vigo, 36310 Pontevedra, Spain*

ARTICLE INFO

Keywords:
Machine learning
Supervised learning
Twitter
Financial stock market
Feature selection

ABSTRACT

Social media plays an important role in facilitating conversations and news dissemination. Specifically, Twitter has recently seen use by investors to facilitate discussions surrounding stock exchange-listed companies. Investors depend on timely, credible information being made available in order to make well-informed investment decisions, with credibility being defined as the believability of information. Much work has been done on assessing credibility on Twitter in domains such as politics and natural disaster events, but the work on assessing the credibility of financial statements is scant within the literature. Investments made on apocryphal information could hamper efforts of social media's aim of providing a transparent arena for sharing news and encouraging discussion of stock market events. This paper presents a novel methodology to assess the credibility of financial stock market tweets, which is evaluated by conducting an experiment using tweets pertaining to companies listed on the London Stock Exchange. Three sets of traditional machine learning classifiers (using three different feature sets) are trained using an annotated dataset. We highlight the importance of considering features specific to the domain in which credibility needs to be assessed for – in the case of this paper, financial features. In total, after discarding non-informative features, 34 general features are combined with over 15 novel financial features for training classifiers. Results show that classifiers trained on both general and financial features can yield improved performance than classifiers trained on general features alone, with Random Forest being the top performer, although the Random Forest model requires more features (37) than that of other classifiers (such as K-Nearest Neighbours – 9) to achieve such performance.

1. Introduction

Investments made on stock markets depend on timely and credible information being made available to investors. Twitter has seen increased use in recent years as a means of sharing information relating to companies listed on stock exchanges (Ranco et al., 2015). The time-critical nature of investing means that investors need to be confident that the news they are consuming is credible and trustworthy. Credibility is generally defined as the believability of information (Sujoy Sikdar, Kang, O'donovan, Höllerer, & Adal, 2013), with social media credibility defined as the aspect of information credibility that can be assessed using only the information available in a social media platform (Castillo et al., 2011). People judge the credibility of general statements based on different constructs such as objectiveness, accuracy, timeliness and reliability (Sujoy Sikdar, Kang, O'donovan, & Höllerer, 2013). Specifically, in terms of Twitter, tweet content and metadata (referred to as features herein), such as the number of followers a user has, and how long they have been a member of Twitter have been seen as informative features for determining the credibility of both the content of the tweet, and the user posting it (de Marcellis-Warin et al., 2017). The problem with such features (namely a user's follower count) is that they can be artificially inflated, as users can obtain thousands of followers from Twitter follower markets within minutes (Stringhini et al., 2013), giving a false indication that the user has a large follower base and is credible (De Micheli & Stroppa, 2013). Determining the credibility of a tweet which is financial in nature becomes even more challenging due to the regulators and exchanges need to quickly curb the spread of misinformation surrounding stocks. Specifically, Twitter users seeking to capitalize on news surrounding stocks by leveraging Twitter's trademark fast information dissemination may be susceptible to rumours and acting upon incredible information within tweets (Da Cruz & De Filgueiras Gomes, 2013). Recent research has found that Twitter is becoming a hotbed for rumour propagation (Maddock et al., 2015). Although such rumours and speculation on Twitter can be informative, as this can reflect investor mood and outlook (Ceccarelli et al., 2016), this new age of financial media in which discussions take place on social media

[*] Corresponding author.
E-mail address: l.evans@mmu.ac.uk (L. Evans).

https://doi.org/10.1016/j.eswa.2020.114351
Received 16 March 2020; Received in revised form 28 August 2020; Accepted 17 November 2020
Available online 20 November 2020

demands mechanisms to assess the credibility of such posts. Repercussions for investors include being cajoled into investing based on apocryphal or incredible information and losing confidence in using a platform such as Twitter if such a platform can be used by perfidious individuals with impunity (De Franco et al., 2007). Twitter does not just act as a discussion board for the investor community, but also acts as an aggregator of financial information by companies and regulators. The financial investment community is currently bereft of ways to assess the credibility of financial stock tweets, as previous work in this field has focused primarily on specific areas such as politics and natural disaster events (Alrubaian et al., 2018).

To this end, one must define what constitutes a financial stock tweet and what is meant by determining the credibility of a financial stock tweet. This paper defines a financial stock tweet as any tweet which contains an occurrence of a stock exchange-listed company's ticker symbol, pre-fixed with a dollar symbol, referred to as a cashtag within the Twitter community. Twitter's cashtag mechanism has been utilised by several works for the purposes of collecting and analysing stock discussion (Oliveira et al., 2016, 2017; Cresci et al., 2018). Although tweets may be relating to a financial stock discussion and not contain a cashtag, this paper takes the stance that tweets are more likely to be related to stock discussions if cashtags are present, and this research focuses on such tweets. We define the credibility of a financial stock tweet as being three-fold: (1) is the cashtag(s) within the tweet related to a specific exchange-listed company? (2) how credible (based on the definition above) is the information within the tweet? and (3) how credible is the author circulating the information? We adopt the definition of user credibility from past research as being the user's perceived trustworthiness and expertise (Liu et al., 2012).

The main contribution of this paper is a novel methodology for assessing the credibility of financial stock tweets on Twitter. The methodology is based on feature extraction and selection according to the relevance of the different features according to an annotated training set. We propose a rich set of features divided into two groups – general features found in all tweets, regardless of subject matter, and financial features, which are engineered specifically to assess the credibility of financial stock tweets. We train three different sets of traditional machine learning classifiers, (1) trained on the *general* features, (2) trained on the *financial* features, and (3) trained on both general and financial feature sets – to ascertain if financial features provide added value in assessing the credibility of financial stock tweets. The methodology proposed in this paper is a generalizable approach which can be applied to any stock exchange, with a slight customisation of the financial features proposed depending on the stock exchange. An experiment utilising tweets pertaining to companies listed on the London Stock Exchange is presented in this paper to validate the proposed financial credibility methodology. The motivation of this paper is to highlight the importance of incorporating features from the domain in which one wishes to assess the credibility of tweets for. The novelty of this work lies in the incorporation of financial features for assessing the credibility of tweets relating to the discussion of stocks.

The research questions this paper will address are as follows:

RQ 1: Can features found in any tweet, regardless of subject matter (i.e. general features), provide an accurate measure for credibility classification of the tweet?

RQ 2: Can financial features, engineered with the intent of assessing the financial credibility of a stock tweet, provide improved classification performance (over the general features) when combined with the general features?

In addition to the methodology for assessing the financial credibility of stock tweets, the other key contributions of this paper can be summarised as follows:

- We present a novel set of financial features for the purpose of assessing the financial credibility of stock tweets
- We highlight the importance of performing feature selection for assessing financial credibility of stock tweets, particularly for machine learning models which do not have inherent feature selection mechanisms embedded within them.

The remainder of this paper is organised as follows: Section 2 explores the related work on the credibility of microblog posts. Section 3 provides an overview of the methodology used. Section 4 outlines the proposed features used to train the machine learning models. Section 5 describes the feature selection techniques used within the methodology. Section 6 outlines the experimental design used to validate the methodology. Section 7 provides a discussion of the results obtained. Section 8 concludes the work undertaken and outlines avenues of potential future work.

2. Background

Although there has been no research on the credibility of financial stock-related tweets, work does exist on the credibility of tweets in areas such as politics (Sujoy Sikdar, Kang, O'donovan, Höllerer, & Adal, 2013; Page & Duffy, 2018), health (Bhattacharya et al., 2012), and natural disaster events (Yang et al., 2019; Thomson et al., 2012). Although some work has been undertaken on determining credibility based on unsupervised approaches (Alrubaian et al., 2018), the related work on credibility assessment is comprised mainly of supervised approaches, which we now explore.

2.1. *Tweet credibility*

The majority of studies of credibility assessment on Twitter are comprised of supervised approaches, predominately decision trees, support vector machines, and Bayesian algorithms (Alrubaian et al., 2018). An extensive survey into the work of credibility on Twitter has been undertaken by Alrubaian et al. (2018), in which they looked at 112 papers on the subject of microblog credibility over the period 2006–2017. Alrubaian et al. (2018) cited one of the key challenges of credibility assessment is that there is a great deal of literature which has developed different credibility dimensions and definitions and that a unified definition of what constitutes credible information does not exist. This section will now explore the related work on supervised learning approaches for determining credibility, due to its popularity versus unsupervised approaches.

Castillo et al. (2011) were amongst the first to undertake research on the credibility of tweets, this work involved assessing the credibility of current news events during a two-month window. Their approach, which made use of Naïve Bayes, Logistic Regression, and Support Vector Machine, was able to correctly recognize 89% of topic appearances and their credibility classification achieved precision and recall scores in the range of 70–80%. Much of the work undertaken since has built upon the initial features proposed in this work. Morris et al. (2012) conducted a series of experiments which included identifying features which are highly relevant for assessing credibility. Their initial experiment found that there are several key features for assessing credibility, which include predominately user-based features such as the author's expertise of the particular topic being assessed (as judged by the author's profile description) and the user's reputation (verified account symbol). In a secondary experiment, they found that the topics of the messages influenced the perception of tweet credibility, with topics in the field of science receiving a higher rating, followed by politics and entertainment. Although the authors initially found that user images had no significant impact on tweet credibility, a follow-up experiment did establish that users who possess the default Twitter icon as their profile picture lowered credibility perception (Morris et al., 2012). Features which are derived from the author of the tweet have been studied intently within the literature, such features derived from the user have been criticised in recent works (Alrubaian et al., 2018)(Stringhini et al.,

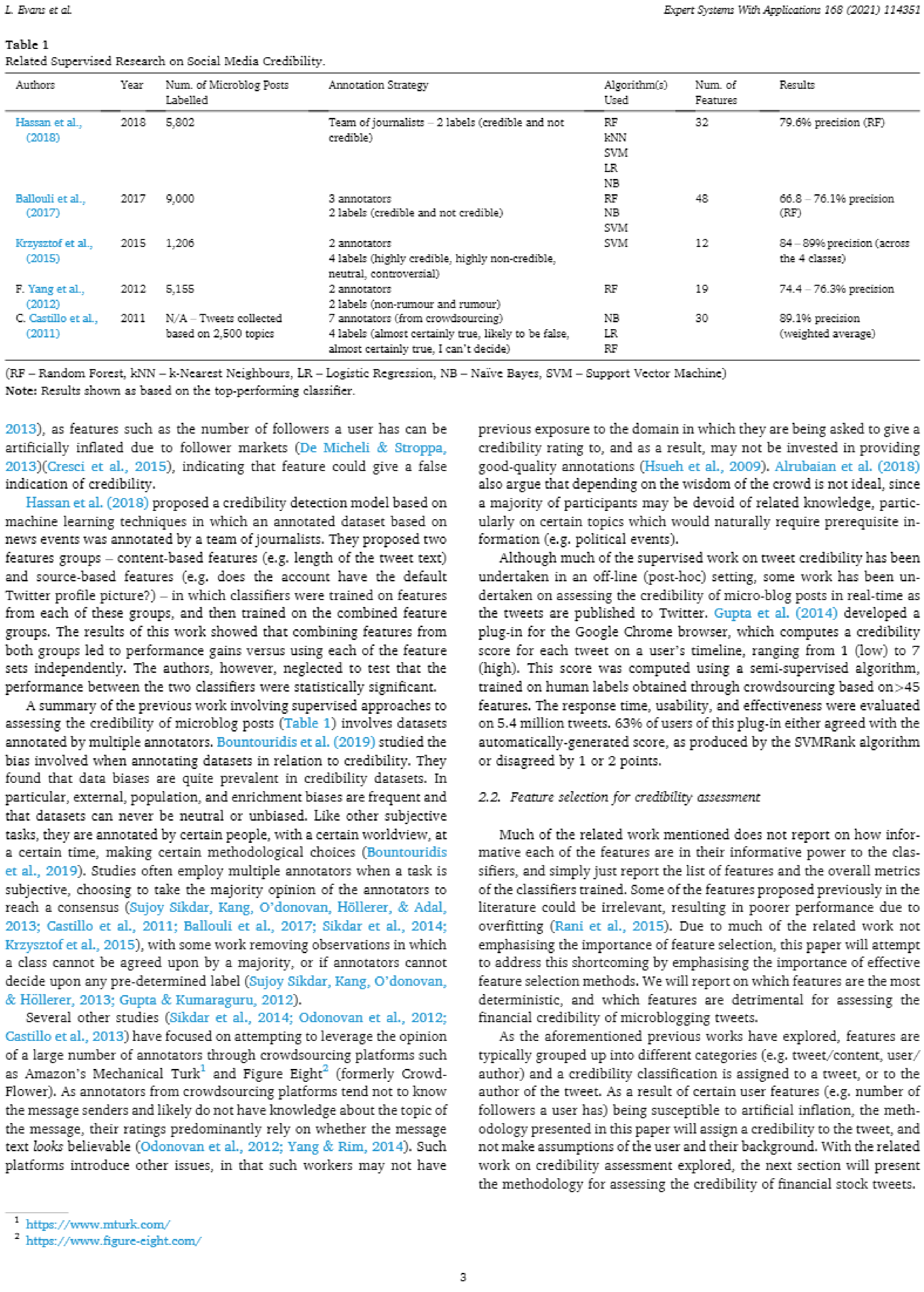


Table 1
Related Supervised Research on Social Media Credibility.

Authors	Year	Num. of Microblog Posts Labelled	Annotation Strategy	Algorithm(s) Used	Num. of Features	Results
Hassan et al., (2018)	2018	5,802	Team of journalists – 2 labels (credible and not credible)	RF kNN SVM LR NB	32	79.6% precision (RF)
Ballouli et al., (2017)	2017	9,000	3 annotators 2 labels (credible and not credible)	RF NB SVM	48	66.8 – 76.1% precision (RF)
Krzysztof et al., (2015)	2015	1,206	2 annotators 4 labels (highly credible, highly non-credible, neutral, controversial)	SVM	12	84 – 89% precision (across the 4 classes)
F. Yang et al., (2012)	2012	5,155	2 annotators 2 labels (non-rumour and rumour)	RF	19	74.4 – 76.3% precision
C. Castillo et al., (2011)	2011	N/A – Tweets collected based on 2,500 topics	7 annotators (from crowdsourcing) 4 labels (almost certainly true, likely to be false, almost certainly true, I can't decide)	NB LR RF	30	89.1% precision (weighted average)

(RF – Random Forest, kNN – k-Nearest Neighbours, LR – Logistic Regression, NB – Naïve Bayes, SVM – Support Vector Machine)
Note: Results shown as based on the top-performing classifier.

2013), as features such as the number of followers a user has can be artificially inflated due to follower markets (De Micheli & Stroppa, 2013)(Cresci et al., 2015), indicating that feature could give a false indication of credibility.

Hassan et al. (2018) proposed a credibility detection model based on machine learning techniques in which an annotated dataset based on news events was annotated by a team of journalists. They proposed two features groups – content-based features (e.g. length of the tweet text) and source-based features (e.g. does the account have the default Twitter profile picture?) – in which classifiers were trained on features from each of these groups, and then trained on the combined feature groups. The results of this work showed that combining features from both groups led to performance gains versus using each of the feature sets independently. The authors, however, neglected to test that the performance between the two classifiers were statistically significant.

A summary of the previous work involving supervised approaches to assessing the credibility of microblog posts (Table 1) involves datasets annotated by multiple annotators. Bountouridis et al. (2019) studied the bias involved when annotating datasets in relation to credibility. They found that data biases are quite prevalent in credibility datasets. In particular, external, population, and enrichment biases are frequent and that datasets can never be neutral or unbiased. Like other subjective tasks, they are annotated by certain people, with a certain worldview, at a certain time, making certain methodological choices (Bountouridis et al., 2019). Studies often employ multiple annotators when a task is subjective, choosing to take the majority opinion of the annotators to reach a consensus (Sujoy Sikdar, Kang, O'donovan, Höllerer, & Adal, 2013; Castillo et al., 2011; Ballouli et al., 2017; Sikdar et al., 2014; Krzysztof et al., 2015), with some work removing observations in which a class cannot be agreed upon by a majority, or if annotators cannot decide upon any pre-determined label (Sujoy Sikdar, Kang, O'donovan, & Höllerer, 2013; Gupta & Kumaraguru, 2012).

Several other studies (Sikdar et al., 2014; Odonovan et al., 2012; Castillo et al., 2013) have focused on attempting to leverage the opinion of a large number of annotators through crowdsourcing platforms such as Amazon's Mechanical Turk[1] and Figure Eight[2] (formerly Crowd-Flower). As annotators from crowdsourcing platforms tend not to know the message senders and likely do not have knowledge about the topic of the message, their ratings predominantly rely on whether the message text *looks* believable (Odonovan et al., 2012; Yang & Rim, 2014). Such platforms introduce other issues, in that such workers may not have previous exposure to the domain in which they are being asked to give a credibility rating to, and as a result, may not be invested in providing good-quality annotations (Hsueh et al., 2009). Alrubaian et al. (2018) also argue that depending on the wisdom of the crowd is not ideal, since a majority of participants may be devoid of related knowledge, particularly on certain topics which would naturally require prerequisite information (e.g. political events).

Although much of the supervised work on tweet credibility has been undertaken in an off-line (post-hoc) setting, some work has been undertaken on assessing the credibility of micro-blog posts in real-time as the tweets are published to Twitter. Gupta et al. (2014) developed a plug-in for the Google Chrome browser, which computes a credibility score for each tweet on a user's timeline, ranging from 1 (low) to 7 (high). This score was computed using a semi-supervised algorithm, trained on human labels obtained through crowdsourcing based on>45 features. The response time, usability, and effectiveness were evaluated on 5.4 million tweets. 63% of users of this plug-in either agreed with the automatically-generated score, as produced by the SVMRank algorithm or disagreed by 1 or 2 points.

2.2. *Feature selection for credibility assessment*

Much of the related work mentioned does not report on how informative each of the features are in their informative power to the classifiers, and simply just report the list of features and the overall metrics of the classifiers trained. Some of the features proposed previously in the literature could be irrelevant, resulting in poorer performance due to overfitting (Rani et al., 2015). Due to much of the related work not emphasising the importance of feature selection, this paper will attempt to address this shortcoming by emphasising the importance of effective feature selection methods. We will report on which features are the most deterministic, and which features are detrimental for assessing the financial credibility of microblogging tweets.

As the aforementioned previous works have explored, features are typically grouped up into different categories (e.g. tweet/content, user/author) and a credibility classification is assigned to a tweet, or to the author of the tweet. As a result of certain user features (e.g. number of followers a user has) being susceptible to artificial inflation, the methodology presented in this paper will assign a credibility to the tweet, and not make assumptions of the user and their background. With the related work on credibility assessment explored, the next section will present the methodology for assessing the credibility of financial stock tweets.

[1] https://www.mturk.com/
[2] https://www.figure-eight.com/

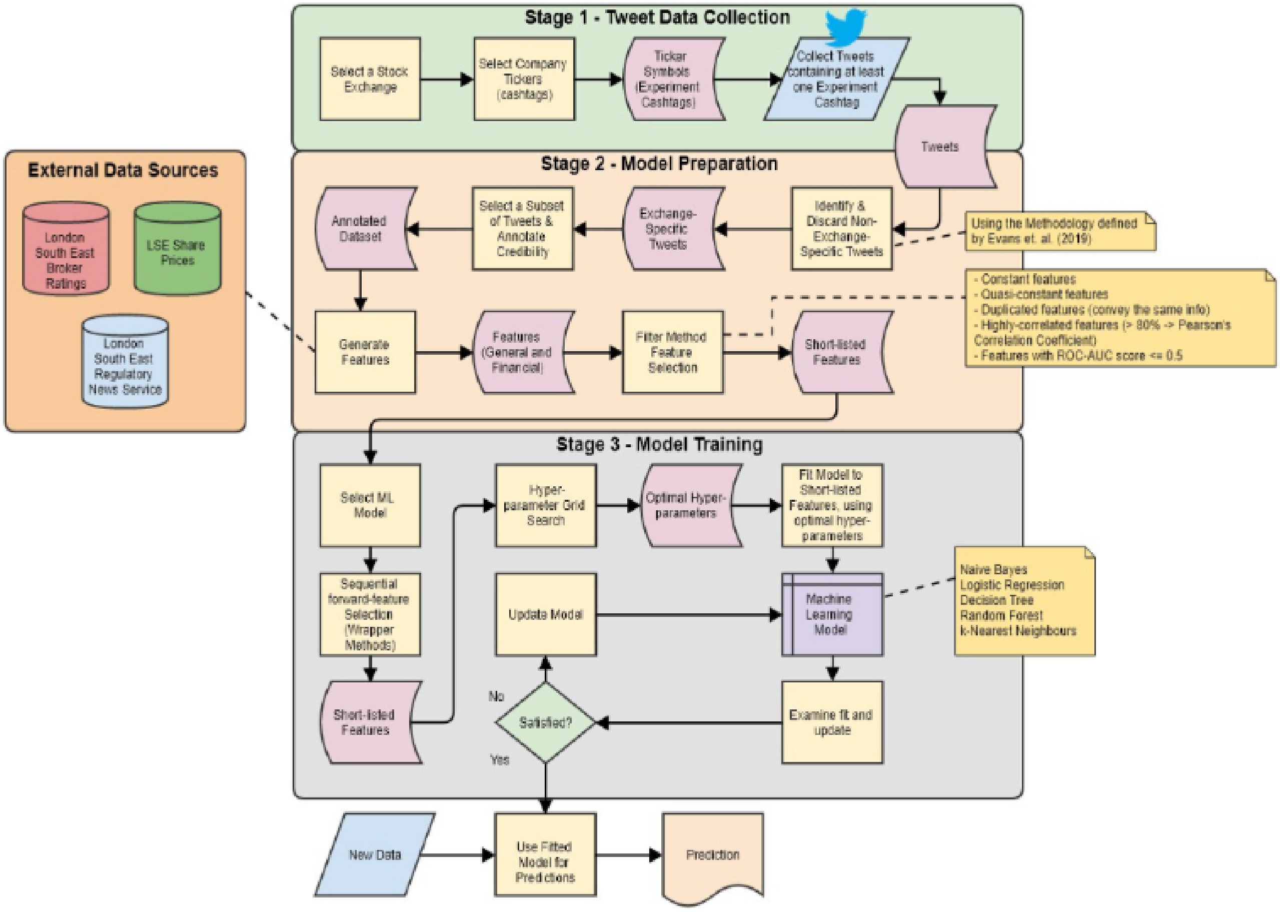

Stage 1 - Tweet Data Collection
Select a Stock Exchange
Select Company Tickers (cashtags)
Ticker Symbols (Experiment Cashtags)
Collect Tweets containing at least one Experiment Cashtag
Tweets
External Data Sources
London South East Broker Ratings
LSE Share Prices
London South East Regulatory News Service
Stage 2 - Model Preparation
Annotated Dataset
Select a Subset of Tweets & Annotate Credibility
Exchange-Specific Tweets
Identify & Discard Non-Exchange-Specific Tweets
Using the Methodology defined by Evans et. al. (2019)
Generate Features
Features (General and Financial)
Filter Method Feature Selection
Short-listed Features
- Constant features
- Quasi-constant features
- Duplicated features (convey the same info)
- Highly-correlated features (> 80% -> Pearson's Correlation Coefficient)
- Features with ROC-AUC score <= 0.5
Stage 3 - Model Training
Select ML Model
Hyper-parameter Grid Search
Optimal Hyper-parameters
Fit Model to Short-listed Features, using optimal hyper-parameters
Sequential forward-feature Selection (Wrapper Methods)
Update Model
Machine Learning Model
Naive Bayes
Logistic Regression
Decision Tree
Random Forest
k-Nearest Neighbours
Short-listed Features
No
Satisfied?
Yes
Examine fit and update
New Data
Use Fitted Model for Predictions
Prediction

Fig. 1. Financial Credibility Assessment Methodology.

3. Methodology

Motivated by the success of supervised learning approaches in assessing the credibility of microblogging posts, we propose a methodology (Fig. 1) to assess the credibility of financial stock tweets (based on our definition of a stock tweet in Section 1). The methodology is comprised of three stages – the first stage of the methodology involves selecting a stock exchange in which to assess the credibility of financial stock tweets. With a stock exchange selected, a list of companies, and their associated ticker symbols can then be shortlisted in which to collect tweets. The second stage involves preparing the data for training machine learning classifiers by performing various feature selection techniques, explained in detail in Section 5. The final stage is the model training stage, in which models are trained on different feature groups with their respective performances being compared to ascertain if the proposed financial features result in more accurate machine learning models. This methodology will be validated by an experiment tailored for a specific stock exchange, explained further in Section 6. We now explain the motivation for each of these stages below.

3.1. Stage 1 – Data collection

The first step of the data collection stage is to select a stock exchange in which to collect stock tweets. Companies are often simultaneously listed on multiple exchanges worldwide (Gregoriou, 2015), meaning statements made about a specific exchange-listed company's share price may not be applicable to the entire company's operations. A shortlist of company ticker symbols can then be created to collect tweets for. Tweets can be collected through the official Twitter API (specific details discussed in Section 6.2). Once tweets have been collected for a given period for a shortlisted list of company ticker symbols (cashtags), tweets can be further analysed to determine if the tweet is associated with a stock-exchange listed company – the primary goal of the second stage of the methodology – discussed next.

3.2. Stage 2 – Model preparation

The second stage is primarily concerned with selecting and generating the features required to train the machine learning classifiers (Section 4) and to perform a quick screening of the features to identify those which are non-informative (e.g. due to being constant or highly-correlated with other features). Before any features can be generated,

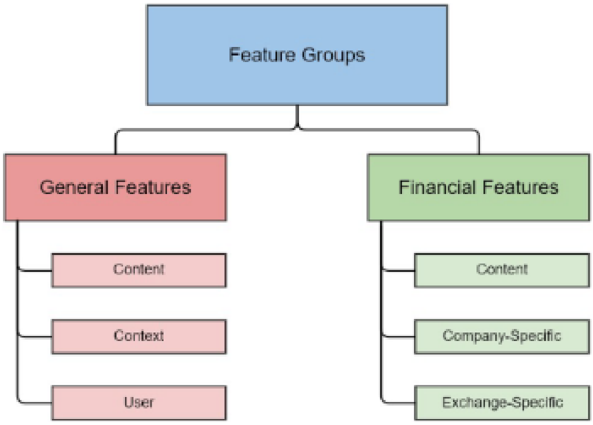


Fig. 2. Feature Subgroups.

however, it is important to note that identifying and collecting tweets for companies for a specific exchange is not always a straightforward task, as we will now discuss in the next subsection.

3.2.1. *Identification of stock exchange-specific tweets*

The primary issue of collecting financial tweets is that any user can create their own cashtag simply by prefixing any word with a dollar symbol ($). As cashtags mimic the company's ticker symbol, companies with identical symbols listed on different stock exchanges share the same cashtag (e.g. $TSCO refers to Tesco PLC on the London Stock Exchange, but also the Tractor Supply Company on the NASDAQ). This has been referred to as a cashtag collision within the literature, with previous work (Evans et al., 2019) adopting trained classifiers to resolve such collisions so that exchange-specific tweets can be identified, and non-stock-related market tweets can be discarded. We utilise the methodology of (Evans et al., 2019) to ensure the collection of exchange-specific tweets and is considered a data cleaning step. Once a suitable subsample of tweets has been obtained after discarding tweets not relating to the pre-chosen exchange, features can then be generated for each of the observations.

3.2.2. *Dataset annotation*

As supervised machine learning models are to be trained, a corpus of tweets must be annotated based on a pre-defined labelled system. As discussed in the related work on supervised learning approaching for credibility assessment (Section 2.1), this is sometimes approached as a binary classification problem (i.e. the tweet is either credible or not credible), with some work opting for more granularity of labels by incorporating labels to indicate the tweet does not have enough information to provide a label in either direction. Section 6.3 includes a detailed overview of the annotation process undertaken for the experiment within this paper.

3.2.3. *Feature engineering and selection*

After an annotated dataset has been obtained, the features can be analysed through appropriate filter-based feature selection techniques in an attempt to reduce the feature space, which may result in more robust machine learning models (Rong et al., 2019). Such filter methods include identifying constant or quasi-constant features, duplicated features which convey the same information, and features which are highly correlated with one another (Bommert et al., 2020). Section 5 provides a detailed overview of each of the feature methods in this work.

3.3. *Stage 3 – Model training*

The final stage of the methodology involves further feature selection techniques (discussed in Section 5) through repeated training of classifiers to discern optimal feature sets by adopting techniques such as wrapper methods. Once an optimal feature subset has been identified, the methodology proposes performing a hyperparameter grid search to further improve the performance of the various classifiers. Although the methodology proposes training traditional supervised classifiers, this list is not exhaustive and can be adapted to include other supervised approaches. The next section introduces the proposed general and financial features to train the machine learning models.

4. Proposed features

Many of the general features (GF) we propose have been used in previous work on the assessment of tweet credibility (Alrubaian et al., 2018). The full list of proposed features (both general and financial), along with a description of each feature can be found in Appendix A. We concede that not every feature proposed will offer an equal amount of informative power to a classification model, and as a result, we do not attempt to justify each of the features in turn, but instead remove the feature(s) if they are found to be of no informative value to the classifiers. The general and financial feature groups, including their associated sub-groups, are provided in Fig. 2.

4.1. *General features (GF)*

The GF group is divided into three sub-groups – content, context, and user. Content features are derived from the viewable content of the tweet. Context features are concerned with information relating to *how* the tweet was created, including the date and time and source of the tweet. User features are concerned with the author of the tweet. Each of these sub-groups will now be discussed further.

4.1.1. *Content*

Content-derived features are features directly accessible from the tweet text or can be engineered from the tweet text. The features proposed in this group include the count of different keyword groups (e.g. noun, verb) and details of the URLs found within the tweet. Many of the features within this group assists in the second dimension of financial tweet credibility – how credible is the information within the tweet?

4.1.2. *Context*

Features within the context sub-group include when the tweet was published to Twitter, in addition to extracting the number of live URLs from the tweet. We argue that simply the presence of a URL should not be seen as a sign of credibility, as it could be the case that the URL is not active in the sense it redirects to a web server. The count of live URLs within the tweet (F27 - Table A1) involves visiting each of the URLs in the tweet to establish if the URL is still live. We define a live URL as any URL which returns a successful response code (200). The number of popular URLs within the tweet, as determined by the domain popularity ranking website, moz[3].

Tweets can be published to Twitter in a variety of ways – these can typically be grouped into manual or automatic. Manual publishing methods involve the user manually publishing a tweet to Twitter, whereas automatic tweets are published based on rules and triggers (Castillo et al., 2019), such as a specific time of the day. Many providers exist for the automatic publishing of content to Twitter (Saguna et al., 2012), such as TweetDeck, Hootsuite, IFTTT. The Tweet Source feature is encoded based on which approach was used to publish the tweet, as described in Table A1.

[3] https://moz.com/top500

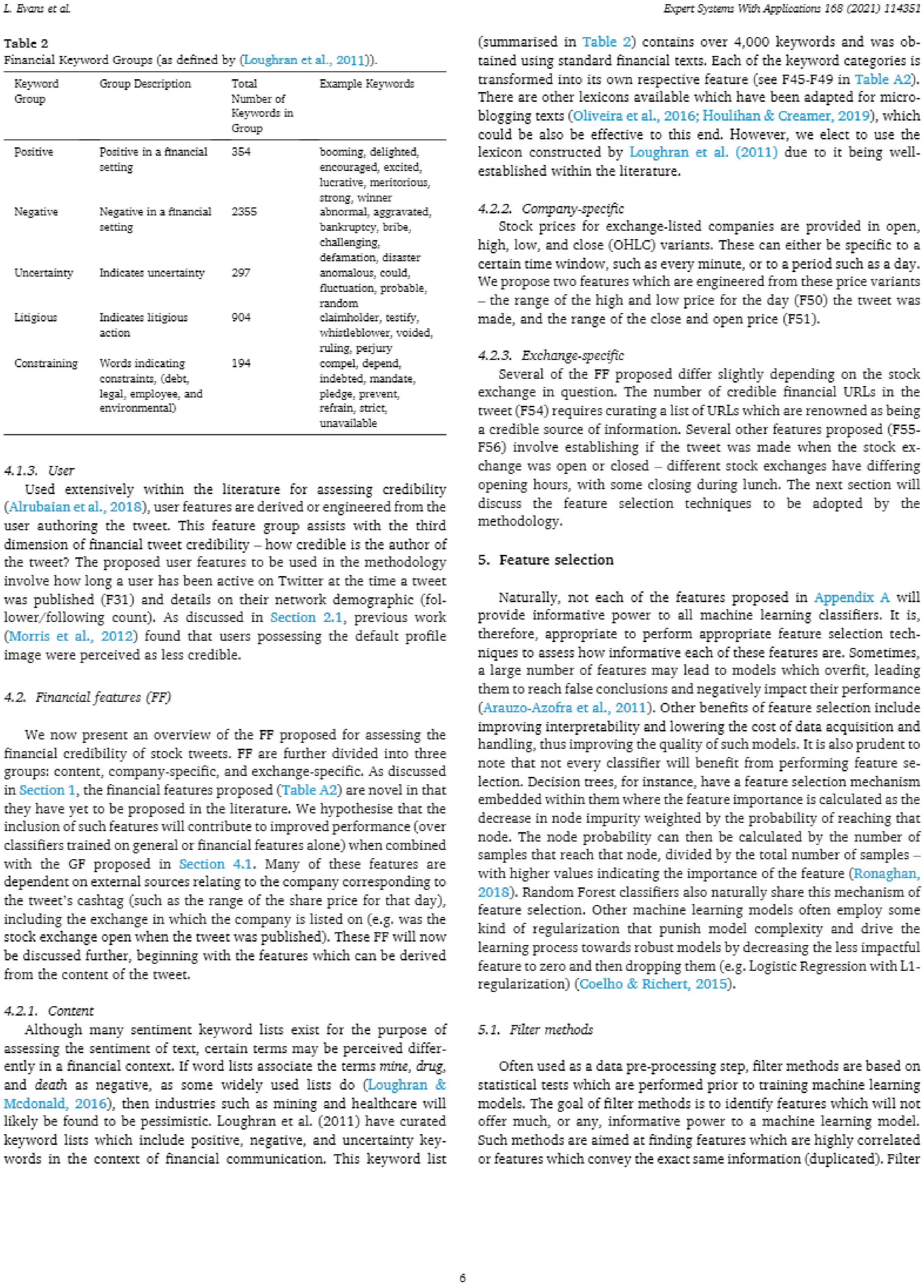


Table 2
Financial Keyword Groups (as defined by (Loughran et al., 2011)).

Keyword Group	Group Description	Total Number of Keywords in Group	Example Keywords
Positive	Positive in a financial setting	354	booming, delighted, encouraged, excited, lucrative, meritorious, strong, winner
Negative	Negative in a financial setting	2355	abnormal, aggravated, bankruptcy, bribe, challenging, defamation, disaster
Uncertainty	Indicates uncertainty	297	anomalous, could, fluctuation, probable, random
Litigious	Indicates litigious action	904	claimholder, testify, whistleblower, voided, ruling, perjury
Constraining	Words indicating constraints, (debt, legal, employee, and environmental)	194	compel, depend, indebted, mandate, pledge, prevent, refrain, strict, unavailable

4.1.3. *User*

Used extensively within the literature for assessing credibility (Alrubaian et al., 2018), user features are derived or engineered from the user authoring the tweet. This feature group assists with the third dimension of financial tweet credibility – how credible is the author of the tweet? The proposed user features to be used in the methodology involve how long a user has been active on Twitter at the time a tweet was published (F31) and details on their network demographic (follower/following count). As discussed in Section 2.1, previous work (Morris et al., 2012) found that users possessing the default profile image were perceived as less credible.

4.2. *Financial features (FF)*

We now present an overview of the **FF** proposed for assessing the financial credibility of stock tweets. **FF** are further divided into three groups: content, company-specific, and exchange-specific. As discussed in Section 1, the financial features proposed (Table A2) are novel in that they have yet to be proposed in the literature. We hypothesise that the inclusion of such features will contribute to improved performance (over classifiers trained on general or financial features alone) when combined with the GF proposed in Section 4.1. Many of these features are dependent on external sources relating to the company corresponding to the tweet's cashtag (such as the range of the share price for that day), including the exchange in which the company is listed on (e.g. was the stock exchange open when the tweet was published). These FF will now be discussed further, beginning with the features which can be derived from the content of the tweet.

4.2.1. *Content*

Although many sentiment keyword lists exist for the purpose of assessing the sentiment of text, certain terms may be perceived differently in a financial context. If word lists associate the terms *mine*, *drug*, and *death* as negative, as some widely used lists do (Loughran & Mcdonald, 2016), then industries such as mining and healthcare will likely be found to be pessimistic. Loughran et al. (2011) have curated keyword lists which include positive, negative, and uncertainty keywords in the context of financial communication. This keyword list (summarised in Table 2) contains over 4,000 keywords and was obtained using standard financial texts. Each of the keyword categories is transformed into its own respective feature (see F45-F49 in Table A2). There are other lexicons available which have been adapted for microblogging texts (Oliveira et al., 2016; Houlihan & Creamer, 2019), which could be also be effective to this end. However, we elect to use the lexicon constructed by Loughran et al. (2011) due to it being well-established within the literature.

4.2.2. *Company-specific*

Stock prices for exchange-listed companies are provided in open, high, low, and close (OHLC) variants. These can either be specific to a certain time window, such as every minute, or to a period such as a day. We propose two features which are engineered from these price variants – the range of the high and low price for the day (F50) the tweet was made, and the range of the close and open price (F51).

4.2.3. *Exchange-specific*

Several of the FF proposed differ slightly depending on the stock exchange in question. The number of credible financial URLs in the tweet (F54) requires curating a list of URLs which are renowned as being a credible source of information. Several other features proposed (F55-F56) involve establishing if the tweet was made when the stock exchange was open or closed – different stock exchanges have differing opening hours, with some closing during lunch. The next section will discuss the feature selection techniques to be adopted by the methodology.

5. Feature selection

Naturally, not each of the features proposed in Appendix A will provide informative power to all machine learning classifiers. It is, therefore, appropriate to perform appropriate feature selection techniques to assess how informative each of these features are. Sometimes, a large number of features may lead to models which overfit, leading them to reach false conclusions and negatively impact their performance (Arauzo-Azofra et al., 2011). Other benefits of feature selection include improving interpretability and lowering the cost of data acquisition and handling, thus improving the quality of such models. It is also prudent to note that not every classifier will benefit from performing feature selection. Decision trees, for instance, have a feature selection mechanism embedded within them where the feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can then be calculated by the number of samples that reach that node, divided by the total number of samples – with higher values indicating the importance of the feature (Ronaghan, 2018). Random Forest classifiers also naturally share this mechanism of feature selection. Other machine learning models often employ some kind of regularization that punish model complexity and drive the learning process towards robust models by decreasing the less impactful feature to zero and then dropping them (e.g. Logistic Regression with L1-regularization) (Coelho & Richert, 2015).

5.1. *Filter methods*

Often used as a data pre-processing step, filter methods are based on statistical tests which are performed prior to training machine learning models. The goal of filter methods is to identify features which will not offer much, or any, informative power to a machine learning model. Such methods are aimed at finding features which are highly correlated or features which convey the exact same information (duplicated). Filter

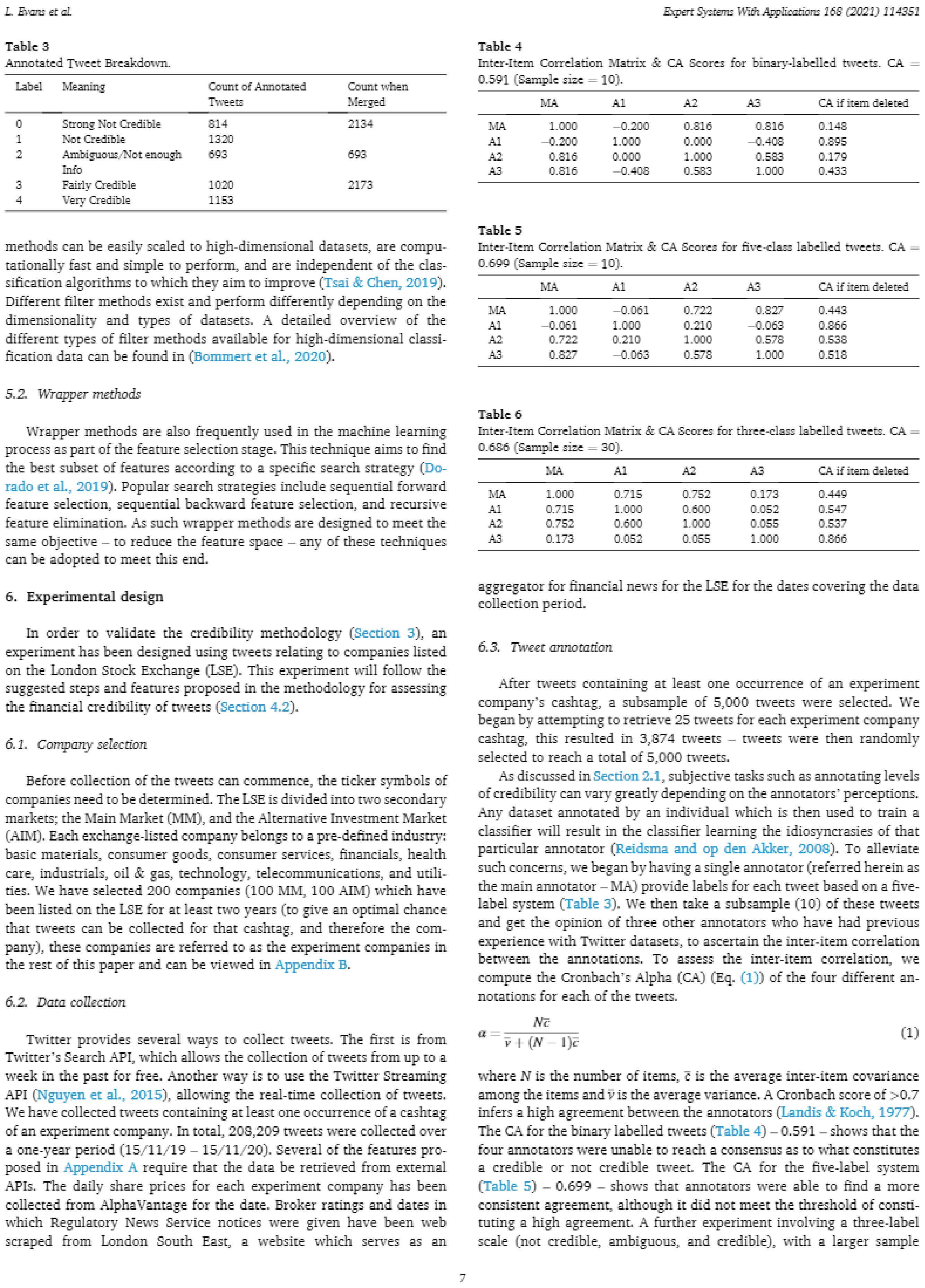


Table 3
Annotated Tweet Breakdown.

Label	Meaning	Count of Annotated Tweets	Count when Merged
0	Strong Not Credible	814	2134
1	Not Credible	1320	
2	Ambiguous/Not enough Info	693	693
3	Fairly Credible	1020	2173
4	Very Credible	1153	

methods can be easily scaled to high-dimensional datasets, are computationally fast and simple to perform, and are independent of the classification algorithms to which they aim to improve (Tsai & Chen, 2019). Different filter methods exist and perform differently depending on the dimensionality and types of datasets. A detailed overview of the different types of filter methods available for high-dimensional classification data can be found in (Bommert et al., 2020).

5.2. *Wrapper methods*

Wrapper methods are also frequently used in the machine learning process as part of the feature selection stage. This technique aims to find the best subset of features according to a specific search strategy (Dorado et al., 2019). Popular search strategies include sequential forward feature selection, sequential backward feature selection, and recursive feature elimination. As such wrapper methods are designed to meet the same objective – to reduce the feature space – any of these techniques can be adopted to meet this end.

6. Experimental design

In order to validate the credibility methodology (Section 3), an experiment has been designed using tweets relating to companies listed on the London Stock Exchange (LSE). This experiment will follow the suggested steps and features proposed in the methodology for assessing the financial credibility of tweets (Section 4.2).

6.1. *Company selection*

Before collection of the tweets can commence, the ticker symbols of companies need to be determined. The LSE is divided into two secondary markets; the Main Market (MM), and the Alternative Investment Market (AIM). Each exchange-listed company belongs to a pre-defined industry: basic materials, consumer goods, consumer services, financials, health care, industrials, oil & gas, technology, telecommunications, and utilities. We have selected 200 companies (100 MM, 100 AIM) which have been listed on the LSE for at least two years (to give an optimal chance that tweets can be collected for that cashtag, and therefore the company), these companies are referred to as the experiment companies in the rest of this paper and can be viewed in Appendix B.

6.2. *Data collection*

Twitter provides several ways to collect tweets. The first is from Twitter's Search API, which allows the collection of tweets from up to a week in the past for free. Another way is to use the Twitter Streaming API (Nguyen et al., 2015), allowing the real-time collection of tweets. We have collected tweets containing at least one occurrence of a cashtag of an experiment company. In total, 208,209 tweets were collected over a one-year period (15/11/19 – 15/11/20). Several of the features proposed in Appendix A require that the data be retrieved from external APIs. The daily share prices for each experiment company has been collected from AlphaVantage for the date. Broker ratings and dates in which Regulatory News Service notices were given have been web scraped from London South East, a website which serves as an aggregator for financial news for the LSE for the dates covering the data collection period.

Table 4
Inter-Item Correlation Matrix & CA Scores for binary-labelled tweets. CA = 0.591 (Sample size = 10).

	MA	A1	A2	A3	CA if item deleted
MA	1.000	−0.200	0.816	0.816	0.148
A1	−0.200	1.000	0.000	−0.408	0.895
A2	0.816	0.000	1.000	0.583	0.179
A3	0.816	−0.408	0.583	1.000	0.433

Table 5
Inter-Item Correlation Matrix & CA Scores for five-class labelled tweets. CA = 0.699 (Sample size = 10).

	MA	A1	A2	A3	CA if item deleted
MA	1.000	−0.061	0.722	0.827	0.443
A1	−0.061	1.000	0.210	−0.063	0.866
A2	0.722	0.210	1.000	0.578	0.538
A3	0.827	−0.063	0.578	1.000	0.518

Table 6
Inter-Item Correlation Matrix & CA Scores for three-class labelled tweets. CA = 0.686 (Sample size = 30).

	MA	A1	A2	A3	CA if item deleted
MA	1.000	0.715	0.752	0.173	0.449
A1	0.715	1.000	0.600	0.052	0.547
A2	0.752	0.600	1.000	0.055	0.537
A3	0.173	0.052	0.055	1.000	0.866

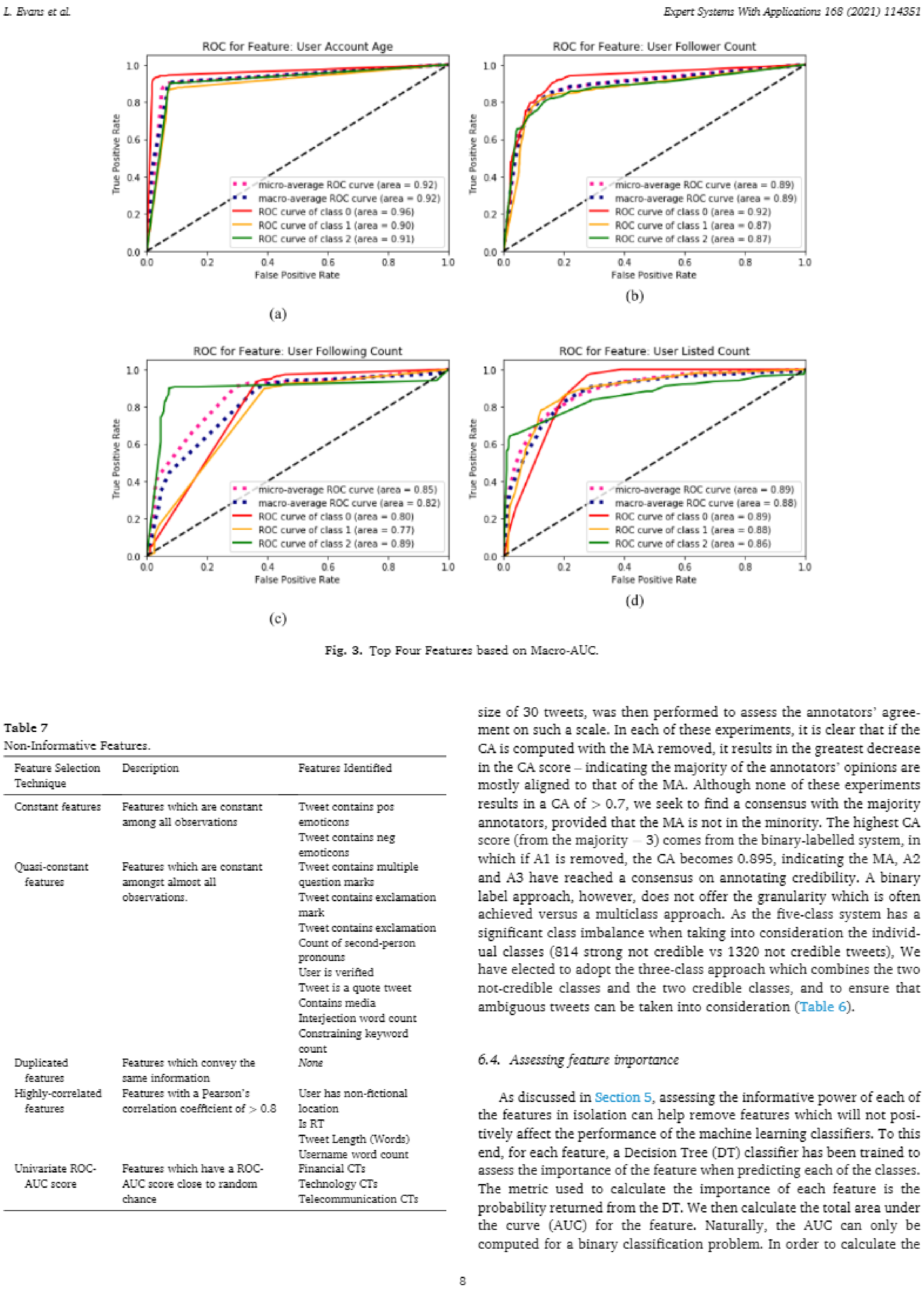
6.3. *Tweet annotation*

After tweets containing at least one occurrence of an experiment company's cashtag, a subsample of 5,000 tweets were selected. We began by attempting to retrieve 25 tweets for each experiment company cashtag, this resulted in 3,874 tweets – tweets were then randomly selected to reach a total of 5,000 tweets.

As discussed in Section 2.1, subjective tasks such as annotating levels of credibility can vary greatly depending on the annotators' perceptions. Any dataset annotated by an individual which is then used to train a classifier will result in the classifier learning the idiosyncrasies of that particular annotator (Reidsma and op den Akker, 2008). To alleviate such concerns, we began by having a single annotator (referred herein as the main annotator – MA) provide labels for each tweet based on a five-label system (Table 3). We then take a subsample (10) of these tweets and get the opinion of three other annotators who have had previous experience with Twitter datasets, to ascertain the inter-item correlation between the annotations. To assess the inter-item correlation, we compute the Cronbach's Alpha (CA) (Eq. (1)) of the four different annotations for each of the tweets.

$$\alpha = \frac{N\bar{c}}{\bar{v} + (N-1)\bar{c}} \qquad (1)$$

where N is the number of items, $\bar{c}$ is the average inter-item covariance among the items and $\bar{v}$ is the average variance. A Cronbach score of >0.7 infers a high agreement between the annotators (Landis & Koch, 1977). The CA for the binary labelled tweets (Table 4) – 0.591 – shows that the four annotators were unable to reach a consensus as to what constitutes a credible or not credible tweet. The CA for the five-label system (Table 5) – 0.699 – shows that annotators were able to find a more consistent agreement, although it did not meet the threshold of constituting a high agreement. A further experiment involving a three-label scale (not credible, ambiguous, and credible), with a larger sample



ROC for Feature: User Account Age
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
micro-average ROC curve (area = 0.92)
macro-average ROC curve (area = 0.92)
ROC curve of class 0 (area = 0.96)
ROC curve of class 1 (area = 0.90)
ROC curve of class 2 (area = 0.91)
ROC for Feature: User Follower Count
micro-average ROC curve (area = 0.89)
macro-average ROC curve (area = 0.89)
ROC curve of class 0 (area = 0.92)
ROC curve of class 1 (area = 0.87)
ROC curve of class 2 (area = 0.87)
ROC for Feature: User Following Count
micro-average ROC curve (area = 0.85)
macro-average ROC curve (area = 0.82)
ROC curve of class 0 (area = 0.80)
ROC curve of class 1 (area = 0.77)
ROC curve of class 2 (area = 0.89)
ROC for Feature: User Listed Count
micro-average ROC curve (area = 0.89)
macro-average ROC curve (area = 0.88)
ROC curve of class 0 (area = 0.89)
ROC curve of class 1 (area = 0.88)
ROC curve of class 2 (area = 0.86)

(a)

(b)

(c)

(d)

Fig. 3. Top Four Features based on Macro-AUC.

Table 7
Non-Informative Features.

Feature Selection Technique	Description	Features Identified
Constant features	Features which are constant among all observations	Tweet contains pos emoticons Tweet contains neg emoticons
Quasi-constant features	Features which are constant amongst almost all observations.	Tweet contains multiple question marks Tweet contains exclamation mark Tweet contains exclamation Count of second-person pronouns User is verified Tweet is a quote tweet Contains media Interjection word count Constraining keyword count
Duplicated features	Features which convey the same information	*None*
Highly-correlated features	Features with a Pearson's correlation coefficient of > 0.8	User has non-fictional location Is RT Tweet Length (Words) Username word count
Univariate ROC-AUC score	Features which have a ROC-AUC score close to random chance	Financial CTs Technology CTs Telecommunication CTs

size of 30 tweets, was then performed to assess the annotators' agreement on such a scale. In each of these experiments, it is clear that if the CA is computed with the MA removed, it results in the greatest decrease in the CA score – indicating the majority of the annotators' opinions are mostly aligned to that of the MA. Although none of these experiments results in a CA of > 0.7, we seek to find a consensus with the majority annotators, provided that the MA is not in the minority. The highest CA score (from the majority = 3) comes from the binary-labelled system, in which if A1 is removed, the CA becomes 0.895, indicating the MA, A2 and A3 have reached a consensus on annotating credibility. A binary label approach, however, does not offer the granularity which is often achieved versus a multiclass approach. As the five-class system has a significant class imbalance when taking into consideration the individual classes (814 strong not credible vs 1320 not credible tweets), We have elected to adopt the three-class approach which combines the two not-credible classes and the two credible classes, and to ensure that ambiguous tweets can be taken into consideration (Table 6).

6.4. *Assessing feature importance*

As discussed in Section 5, assessing the informative power of each of the features in isolation can help remove features which will not positively affect the performance of the machine learning classifiers. To this end, for each feature, a Decision Tree (DT) classifier has been trained to assess the importance of the feature when predicting each of the classes. The metric used to calculate the importance of each feature is the probability returned from the DT. We then calculate the total area under the curve (AUC) for the feature. Naturally, the AUC can only be computed for a binary classification problem. In order to calculate the

Table 8
Classifier Results.

Classifier	General Features						Financial Features						General + Financial Features					
	Features (/34)	Acc	Pre	Rec	F1	AUC	Features (/21)	Acc	Pre	Rec	F1	AUC	Features (/55)	Acc	Pre	Rec	F1	AUC
NB	4	85.5	84.8	85.5	85.0	89.1	12	61.0	63.9	60.3	59.7	70.4	6 (2FF)	85.6	84.9	85.6	85.1	91.4
LR	21	88.0	84.6	86.0	85.3	90.5	9	55.9	40.8	50.7	43.0	64.0	27 (9FF)	87.6	87.1	86.8	86.9	92.0
DT	18	90.1	90.6	90.4	90.5	92.6	10	54.2	55.1	49.6	43.0	63.1	11 (3FF)	89.7	90.1	90.0	90.0	93.1
RF	20	92.7	93.1	92.6	92.9	93.8	11	61.9	63.1	60.9	60.4	70.9	37 (12FF)	93.5	94.3	93.2	93.7	94.3
kNN	7	91.4	92.3	91.1	91.6	93.2	7	61.5	64.0	61.3	60.8	71.1	9 (2FF)	92.7	93.6	92.5	92.9	93.6

Note: Scores presented are the macro average percentage (%).

AUC for a multi-class problem, the DT classifier, which is capable of producing an output $y = \{0, 1, 2\}$, is converted into three binary classifiers through a One-Vs-Rest approach (Ambusaidi et al., 2016). Each of the AUC scores for the three binary classifiers, for each feature, can then be calculated to ascertain the feature's predictive power for each class. The AUC score can be computed in different ways for a multiclass classifier: the macro average computes the metric for each class independently before taking the average, whereas the micro average is the traditional mean for all samples (Aghdam et al., 2009). Macro-averaging treats all classes equally, whereas micro-averaging favours majority classes. We elect to judge the informative power of the feature based on its AUC macro average, due to ambiguous tweets being relatively more uncommon than credible and not credible tweets. Four of the features (Fig. 3) exhibit a macro AUC score of > 0.8, indicating they will likely offer a great degree of informative power when used to train machine learning classifiers. These four features are all contained within the general group and are attributed to the user of the tweet, and is consistent with previous work (Yang et al., 2012) which found that user attributes to be incredibly predictive of credibility.

The filter methods outlined in the methodology (Fig. 1), have been applied to the annotated dataset (5,000 tweets). Based on these five different filter method feature selection techniques, 18 features (Table 7) have been identified to provide no meaningful informative power based on the probability returned from the DT.

With the informative and non-informative features indentified, machine learning classifiers can now be trained on an optimal feature set. The 18 non-informative features identified have been dropped due to the reasons outlined in Table 7.

7. Experimental results & discussion

We now present the results (Table 8) obtained from the experiment based on all of the features after the non-informative features are removed (34 GF, 21 FF), and illustrate that some models' performance suffers if feature selection techniques are not taken into consideration. We have trained classifiers which have demonstrated previous success in assessing the credibility of microblog messages (Naïve Bayes, k-Nearest Neighbours, Decision Trees, Logistic Regression, and Random Forest) (Alrubaian et al., 2018). All of the results obtained are a result of 10-fold cross-validation using an 80/20 train/test split and implemented using the scikit-learn library within Python. Each of the classification models underwent a grid search to find optimal hyperparameters. Three sets of classifiers have been trained; (1) trained on the GF, (2) trained on the FF, and (3) trained on both sets of features.

As indicated by the results of the sequential feature selection (Fig. 4), the kNN and NB classifiers suffer clear decreases in their performance when more features are added to the feature space due to the well-documented phenomenon of the curse of dimensionality (Parmezan et al., 2017). DT, RF, and LR, also suffer minor decreases, although, due to the nature of these three algorithms, they are less impacted. Based on the AUC, the RF classifier is the top-performing classifier when trained on the GF and FF sets respectively. Clearly, classifiers trained solely on the FF pale in performance when compared to classifiers trained on the other feature sets. Regarding RQ1, GF by themselves are extremely informative for assessing the credibility classification of tweets. When combined with FF (RQ2), performance gains are evident in all of the classifiers trained on the combined feature sets. The importance of feature selection is particularly prevalent for the kNN classifier, which reaches its zenith at 9 features and almost outperforms the RF when both are compared at such a feature space size. In terms of which FFs were seen to be informative, the RF trained on the combined features utilised 12 financial features, which included; F46, F55, F56, F58, and 8xF59↓). In respect to the five classifiers trained on the combined features, the most popular FFs utilised by the classifiers were the count of cashtags in the tweet (F58), and the count of technology and healthcare cashtags within the tweet (2xF59↓).

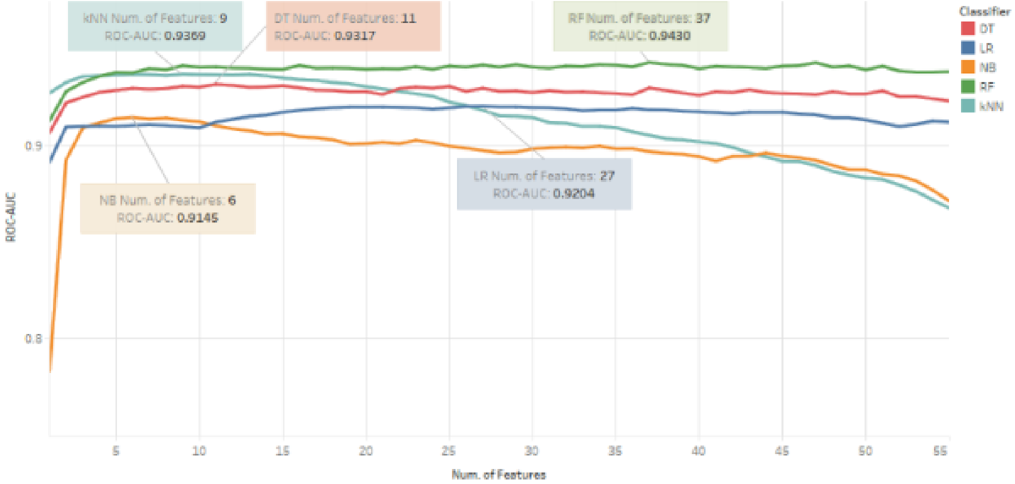


Fig. 4. Sequential Forward Feature Selection Results (Combined features).

As evident from the initial experiment results, RF appears to be the best performing classifier when the feature sets are combined. We now test if the differences between the predictions of the RF trained on GF versus the RF trained on the combined features are statistically significant by conducting the Stuart-Maxwell test. The Stuart-Maxwell test is an extension to the McNemar test, used to assess marginal homogeneity in independent matched-pair data, where responses are allowed more than two response categories (Yang et al., 2011). The p-value of the Stuart-Maxwell test on the predictions of both the RF trained on GF and the RF trained on the combined features is 0.0031, indicating the difference between the two classifiers are statistically significant.

8. Conclusion

This paper has presented a methodology for assessing the credibility of financial stock tweets. Two groups of features were proposed, GF widely used within the literature and a domain-specific group specific to financial stock tweets. Before the training of classifiers, feature selection techniques were used to identify non-informative features. Based on the two groups of features (general and financial), three sets of classifiers were trained, with the first two groups being the set of general and FF respectively, and the third being the combination of the two. Performance gains were noted in the machine learning classifiers, with some classifiers (NB and kNN) suffering when their respective feature spaces grew, undoubtedly due to the curse of dimensionality. Although the RF classifiers were certainly the best performing classifiers in respect to the AUC, it is important to note that the kNN classifier trained on the combined feature set was also a formidable classifier due to its comparative performance with the RF classifiers without having to take into account as many features (9 features compared to 37 for RF). The number of dependent features for the RF classifier presents some limitations for deploying a model dependent on a larger number of features, some of which are more computationally to obtain than others. The count of live URLs within the tweet (F27) requires querying each URL in the tweet, which can be computationally expensive to generate the feature if a tweet contains multiple URLs. Establishing the computational cost of features such as the count of live URLs in a tweet and to assess their suitability in a real-time credibility model is an interesting avenue for future work. There are other features which could be engineered by querying external APIs such as historical stock market values and ascertaining if the tweet contains credible information regarding stock movements of the cashtags contained in the tweet. This would be most beneficial if attempting to classify user credibility – does a user often tweet information about stock-listed companies which turned out to be true? Adopting a lexicon which has been constructed based on financial microblog texts, such as the one constructed by (Oliveira et al., 2016) could yield improved results when assessing tweet credibility, this is an avenue for future work.

As discussed in section 3.3, the list of supervised classifiers in this work is not exhaustive, Support Vector Machines (SVM) were included in the list of classifiers to be trained, but performing hyperparameter grid searches were extremely computationally expensive and were abandoned due to the unsuitability of comparing the SVM classifier with no hyperparameter tuning to that of models which had undergone extensive hyperparameter tuning. Future work in this regard would include the SVM to assess its predictive power in classifying the credibility of financial stock tweets, with neural network architectures also being considered. The credibility methodology presented in this paper will be utilised in the future by a smart data ecosystem, with the intent of monitoring and detecting financial market irregularities.

CRediT authorship contribution statement

Lewis Evans: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Majdi Owda:** Conceptualization, Methodology, Validation, Writing - review & editing, Supervision, Project administration. **Keeley Crockett:** Conceptualization, Methodology, Validation, Writing - review & editing, Supervision, Project administration. **Ana Fernandez Vilas:** Conceptualization, Methodology, Validation, Writing - review & editing, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

Table A1

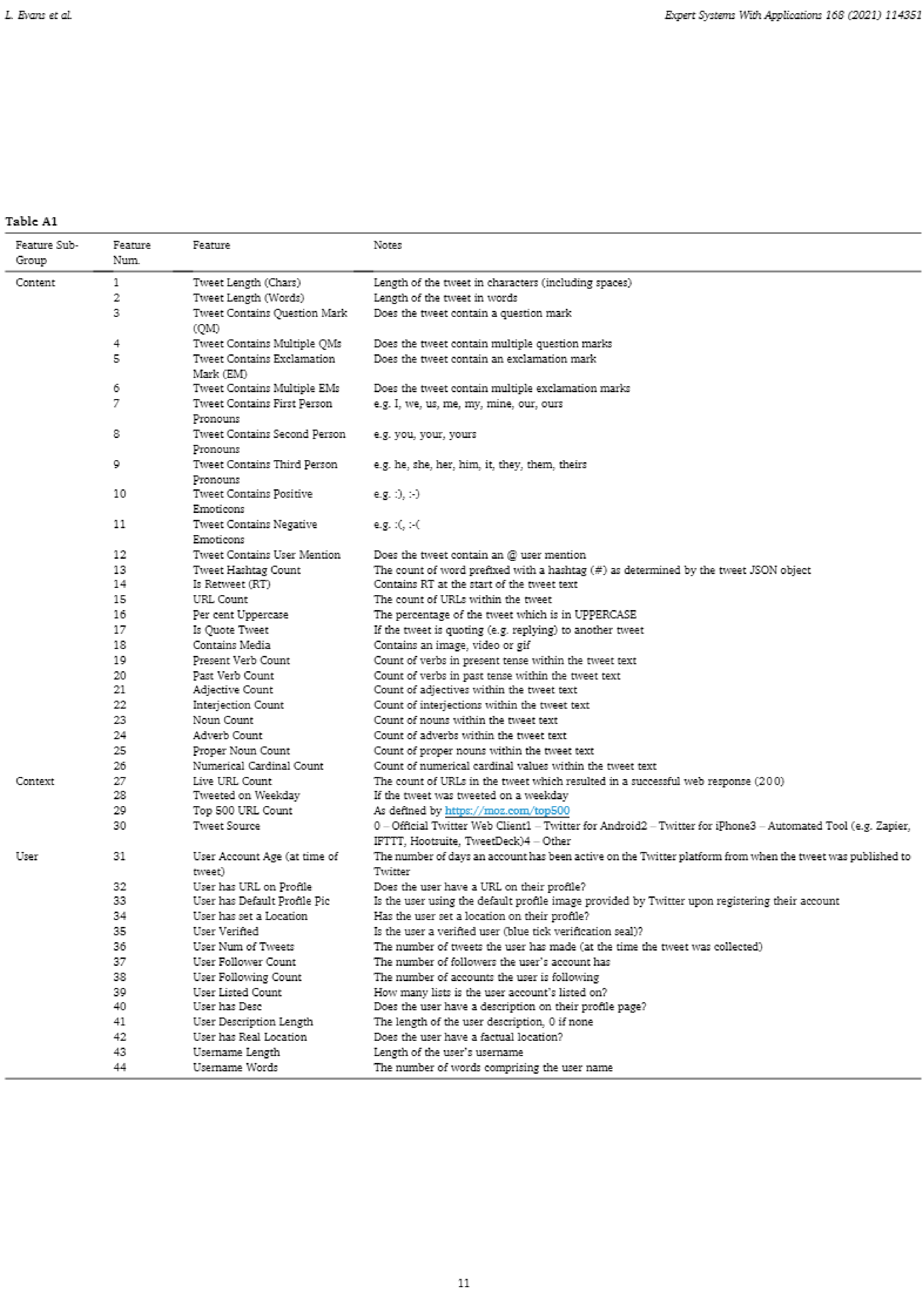


Table A1

Feature Sub-Group	Feature Num.	Feature	Notes
Content	1	Tweet Length (Chars)	Length of the tweet in characters (including spaces)
	2	Tweet Length (Words)	Length of the tweet in words
	3	Tweet Contains Question Mark (QM)	Does the tweet contain a question mark
	4	Tweet Contains Multiple QMs	Does the tweet contain multiple question marks
	5	Tweet Contains Exclamation Mark (EM)	Does the tweet contain an exclamation mark
	6	Tweet Contains Multiple EMs	Does the tweet contain multiple exclamation marks
	7	Tweet Contains First Person Pronouns	e.g. I, we, us, me, my, mine, our, ours
	8	Tweet Contains Second Person Pronouns	e.g. you, your, yours
	9	Tweet Contains Third Person Pronouns	e.g. he, she, her, him, it, they, them, theirs
	10	Tweet Contains Positive Emoticons	e.g. :), :-)
	11	Tweet Contains Negative Emoticons	e.g. :(, :-(
	12	Tweet Contains User Mention	Does the tweet contain an @ user mention
	13	Tweet Hashtag Count	The count of word prefixed with a hashtag (#) as determined by the tweet JSON object
	14	Is Retweet (RT)	Contains RT at the start of the tweet text
	15	URL Count	The count of URLs within the tweet
	16	Per cent Uppercase	The percentage of the tweet which is in UPPERCASE
	17	Is Quote Tweet	If the tweet is quoting (e.g. replying) to another tweet
	18	Contains Media	Contains an image, video or gif
	19	Present Verb Count	Count of verbs in present tense within the tweet text
	20	Past Verb Count	Count of verbs in past tense within the tweet text
	21	Adjective Count	Count of adjectives within the tweet text
	22	Interjection Count	Count of interjections within the tweet text
	23	Noun Count	Count of nouns within the tweet text
	24	Adverb Count	Count of adverbs within the tweet text
	25	Proper Noun Count	Count of proper nouns within the tweet text
	26	Numerical Cardinal Count	Count of numerical cardinal values within the tweet text
Context	27	Live URL Count	The count of URLs in the tweet which resulted in a successful web response (200)
	28	Tweeted on Weekday	If the tweet was tweeted on a weekday
	29	Top 500 URL Count	As defined by https://moz.com/top500
	30	Tweet Source	0 – Official Twitter Web Client1 – Twitter for Android2 – Twitter for iPhone3 – Automated Tool (e.g. Zapier, IFTTT, Hootsuite, TweetDeck)4 – Other
User	31	User Account Age (at time of tweet)	The number of days an account has been active on the Twitter platform from when the tweet was published to Twitter
	32	User has URL on Profile	Does the user have a URL on their profile?
	33	User has Default Profile Pic	Is the user using the default profile image provided by Twitter upon registering their account
	34	User has set a Location	Has the user set a location on their profile?
	35	User Verified	Is the user a verified user (blue tick verification seal)?
	36	User Num of Tweets	The number of tweets the user has made (at the time the tweet was collected)
	37	User Follower Count	The number of followers the user's account has
	38	User Following Count	The number of accounts the user is following
	39	User Listed Count	How many lists is the user account's listed on?
	40	User has Desc	Does the user have a description on their profile page?
	41	User Description Length	The length of the user description, 0 if none
	42	User has Real Location	Does the user have a factual location?
	43	Username Length	Length of the user's username
	44	Username Words	The number of words comprising the user name

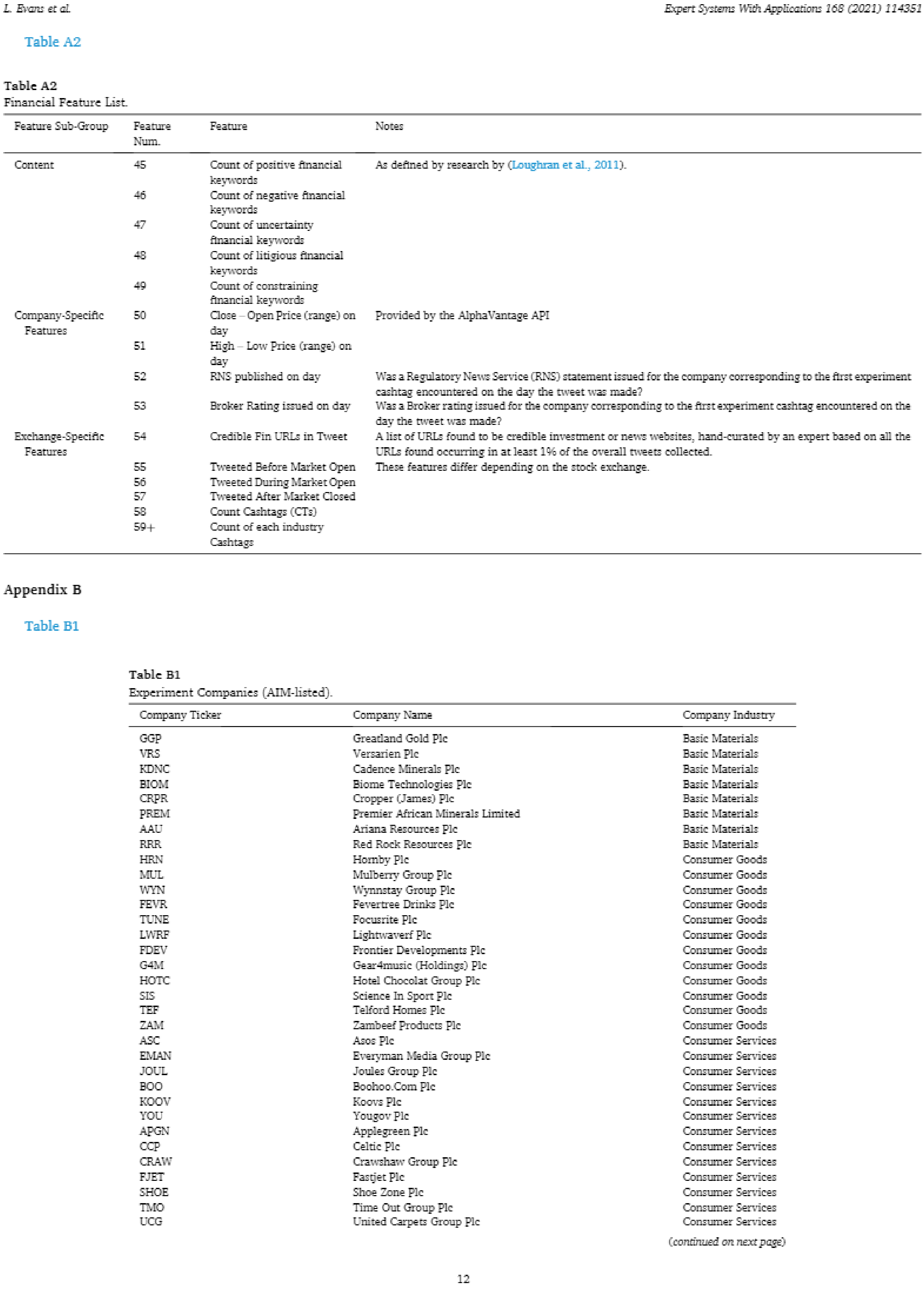


Table A2

Table A2
Financial Feature List.

Feature Sub-Group	Feature Num.	Feature	Notes
Content	45	Count of positive financial keywords	As defined by research by (Loughran et al., 2011).
	46	Count of negative financial keywords	
	47	Count of uncertainty financial keywords	
	48	Count of litigious financial keywords	
	49	Count of constraining financial keywords	
Company-Specific Features	50	Close – Open Price (range) on day	Provided by the AlphaVantage API
	51	High – Low Price (range) on day	
	52	RNS published on day	Was a Regulatory News Service (RNS) statement issued for the company corresponding to the first experiment cashtag encountered on the day the tweet was made?
	53	Broker Rating issued on day	Was a Broker rating issued for the company corresponding to the first experiment cashtag encountered on the day the tweet was made?
Exchange-Specific Features	54	Credible Fin URLs in Tweet	A list of URLs found to be credible investment or news websites, hand-curated by an expert based on all the URLs found occurring in at least 1% of the overall tweets collected.
	55	Tweeted Before Market Open	These features differ depending on the stock exchange.
	56	Tweeted During Market Open	
	57	Tweeted After Market Closed	
	58	Count Cashtags (CTs)	
	59+	Count of each industry Cashtags	

Appendix B

Table B1

Table B1
Experiment Companies (AIM-listed).

Company Ticker	Company Name	Company Industry
GGP	Greatland Gold Plc	Basic Materials
VRS	Versarien Plc	Basic Materials
KDNC	Cadence Minerals Plc	Basic Materials
BIOM	Biome Technologies Plc	Basic Materials
CRPR	Cropper (James) Plc	Basic Materials
PREM	Premier African Minerals Limited	Basic Materials
AAU	Ariana Resources Plc	Basic Materials
RRR	Red Rock Resources Plc	Basic Materials
HRN	Hornby Plc	Consumer Goods
MUL	Mulberry Group Plc	Consumer Goods
WYN	Wynnstay Group Plc	Consumer Goods
FEVR	Fevertree Drinks Plc	Consumer Goods
TUNE	Focusrite Plc	Consumer Goods
LWRF	Lightwaverf Plc	Consumer Goods
FDEV	Frontier Developments Plc	Consumer Goods
G4M	Gear4music (Holdings) Plc	Consumer Goods
HOTC	Hotel Chocolat Group Plc	Consumer Goods
SIS	Science In Sport Plc	Consumer Goods
TEF	Telford Homes Plc	Consumer Goods
ZAM	Zambeef Products Plc	Consumer Goods
ASC	Asos Plc	Consumer Services
EMAN	Everyman Media Group Plc	Consumer Services
JOUL	Joules Group Plc	Consumer Services
BOO	Boohoo.Com Plc	Consumer Services
KOOV	Koovs Plc	Consumer Services
YOU	Yougov Plc	Consumer Services
APGN	Applegreen Plc	Consumer Services
CCP	Celtic Plc	Consumer Services
CRAW	Crawshaw Group Plc	Consumer Services
FJET	Fastjet Plc	Consumer Services
SHOE	Shoe Zone Plc	Consumer Services
TMO	Time Out Group Plc	Consumer Services
UCG	United Carpets Group Plc	Consumer Services

(continued on next page)

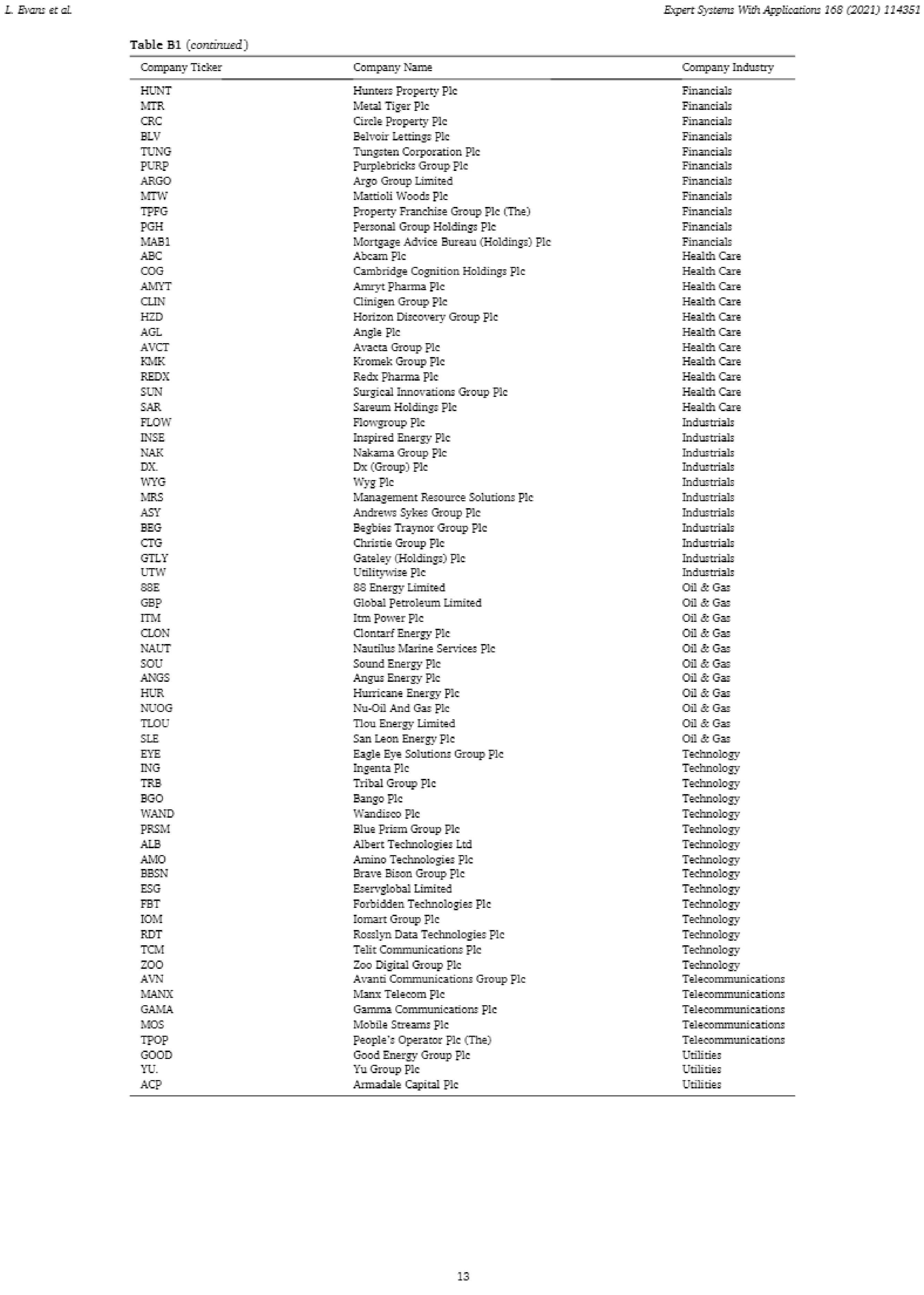


Table B1 (*continued*)

Company Ticker	Company Name	Company Industry
HUNT	Hunters Property Plc	Financials
MTR	Metal Tiger Plc	Financials
CRC	Circle Property Plc	Financials
BLV	Belvoir Lettings Plc	Financials
TUNG	Tungsten Corporation Plc	Financials
PURP	Purplebricks Group Plc	Financials
ARGO	Argo Group Limited	Financials
MTW	Mattioli Woods Plc	Financials
TPFG	Property Franchise Group Plc (The)	Financials
PGH	Personal Group Holdings Plc	Financials
MAB1	Mortgage Advice Bureau (Holdings) Plc	Financials
ABC	Abcam Plc	Health Care
COG	Cambridge Cognition Holdings Plc	Health Care
AMYT	Amryt Pharma Plc	Health Care
CLIN	Clinigen Group Plc	Health Care
HZD	Horizon Discovery Group Plc	Health Care
AGL	Angle Plc	Health Care
AVCT	Avacta Group Plc	Health Care
KMK	Kromek Group Plc	Health Care
REDX	Redx Pharma Plc	Health Care
SUN	Surgical Innovations Group Plc	Health Care
SAR	Sareum Holdings Plc	Health Care
FLOW	Flowgroup Plc	Industrials
INSE	Inspired Energy Plc	Industrials
NAK	Nakama Group Plc	Industrials
DX.	Dx (Group) Plc	Industrials
WYG	Wyg Plc	Industrials
MRS	Management Resource Solutions Plc	Industrials
ASY	Andrews Sykes Group Plc	Industrials
BEG	Begbies Traynor Group Plc	Industrials
CTG	Christie Group Plc	Industrials
GTLY	Gateley (Holdings) Plc	Industrials
UTW	Utilitywise Plc	Industrials
88E	88 Energy Limited	Oil & Gas
GBP	Global Petroleum Limited	Oil & Gas
ITM	Itm Power Plc	Oil & Gas
CLON	Clontarf Energy Plc	Oil & Gas
NAUT	Nautilus Marine Services Plc	Oil & Gas
SOU	Sound Energy Plc	Oil & Gas
ANGS	Angus Energy Plc	Oil & Gas
HUR	Hurricane Energy Plc	Oil & Gas
NUOG	Nu-Oil And Gas Plc	Oil & Gas
TLOU	Tlou Energy Limited	Oil & Gas
SLE	San Leon Energy Plc	Oil & Gas
EYE	Eagle Eye Solutions Group Plc	Technology
ING	Ingenta Plc	Technology
TRB	Tribal Group Plc	Technology
BGO	Bango Plc	Technology
WAND	Wandisco Plc	Technology
PRSM	Blue Prism Group Plc	Technology
ALB	Albert Technologies Ltd	Technology
AMO	Amino Technologies Plc	Technology
BBSN	Brave Bison Group Plc	Technology
ESG	Eservglobal Limited	Technology
FBT	Forbidden Technologies Plc	Technology
IOM	Iomart Group Plc	Technology
RDT	Rosslyn Data Technologies Plc	Technology
TCM	Telit Communications Plc	Technology
ZOO	Zoo Digital Group Plc	Technology
AVN	Avanti Communications Group Plc	Telecommunications
MANX	Manx Telecom Plc	Telecommunications
GAMA	Gamma Communications Plc	Telecommunications
MOS	Mobile Streams Plc	Telecommunications
TPOP	People's Operator Plc (The)	Telecommunications
GOOD	Good Energy Group Plc	Utilities
YU.	Yu Group Plc	Utilities
ACP	Armadale Capital Plc	Utilities

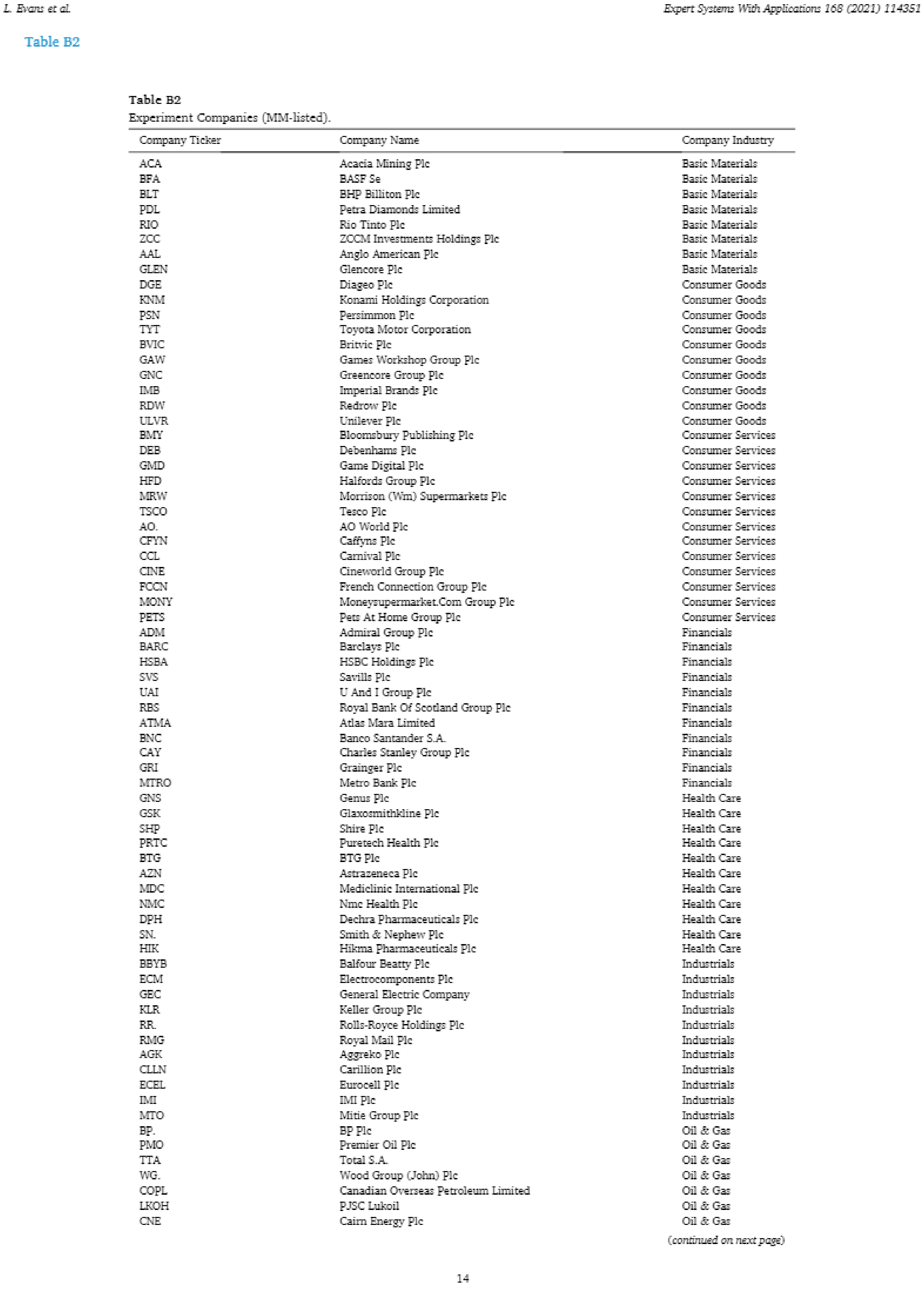


Table B2

Table B2
Experiment Companies (MM-listed).

Company Ticker	Company Name	Company Industry
ACA	Acacia Mining Plc	Basic Materials
BFA	BASF Se	Basic Materials
BLT	BHP Billiton Plc	Basic Materials
PDL	Petra Diamonds Limited	Basic Materials
RIO	Rio Tinto Plc	Basic Materials
ZCC	ZCCM Investments Holdings Plc	Basic Materials
AAL	Anglo American Plc	Basic Materials
GLEN	Glencore Plc	Basic Materials
DGE	Diageo Plc	Consumer Goods
KNM	Konami Holdings Corporation	Consumer Goods
PSN	Persimmon Plc	Consumer Goods
TYT	Toyota Motor Corporation	Consumer Goods
BVIC	Britvic Plc	Consumer Goods
GAW	Games Workshop Group Plc	Consumer Goods
GNC	Greencore Group Plc	Consumer Goods
IMB	Imperial Brands Plc	Consumer Goods
RDW	Redrow Plc	Consumer Goods
ULVR	Unilever Plc	Consumer Goods
BMY	Bloomsbury Publishing Plc	Consumer Services
DEB	Debenhams Plc	Consumer Services
GMD	Game Digital Plc	Consumer Services
HFD	Halfords Group Plc	Consumer Services
MRW	Morrison (Wm) Supermarkets Plc	Consumer Services
TSCO	Tesco Plc	Consumer Services
AO.	AO World Plc	Consumer Services
CFYN	Caffyns Plc	Consumer Services
CCL	Carnival Plc	Consumer Services
CINE	Cineworld Group Plc	Consumer Services
FCCN	French Connection Group Plc	Consumer Services
MONY	Moneysupermarket.Com Group Plc	Consumer Services
PETS	Pets At Home Group Plc	Consumer Services
ADM	Admiral Group Plc	Financials
BARC	Barclays Plc	Financials
HSBA	HSBC Holdings Plc	Financials
SVS	Savills Plc	Financials
UAI	U And I Group Plc	Financials
RBS	Royal Bank Of Scotland Group Plc	Financials
ATMA	Atlas Mara Limited	Financials
BNC	Banco Santander S.A.	Financials
CAY	Charles Stanley Group Plc	Financials
GRI	Grainger Plc	Financials
MTRO	Metro Bank Plc	Financials
GNS	Genus Plc	Health Care
GSK	Glaxosmithkline Plc	Health Care
SHP	Shire Plc	Health Care
PRTC	Puretech Health Plc	Health Care
BTG	BTG Plc	Health Care
AZN	Astrazeneca Plc	Health Care
MDC	Mediclinic International Plc	Health Care
NMC	Nmc Health Plc	Health Care
DPH	Dechra Pharmaceuticals Plc	Health Care
SN.	Smith & Nephew Plc	Health Care
HIK	Hikma Pharmaceuticals Plc	Health Care
BBYB	Balfour Beatty Plc	Industrials
ECM	Electrocomponents Plc	Industrials
GEC	General Electric Company	Industrials
KLR	Keller Group Plc	Industrials
RR.	Rolls-Royce Holdings Plc	Industrials
RMG	Royal Mail Plc	Industrials
AGK	Aggreko Plc	Industrials
CLLN	Carillion Plc	Industrials
ECEL	Eurocell Plc	Industrials
IMI	IMI Plc	Industrials
MTO	Mitie Group Plc	Industrials
BP.	BP Plc	Oil & Gas
PMO	Premier Oil Plc	Oil & Gas
TTA	Total S.A.	Oil & Gas
WG.	Wood Group (John) Plc	Oil & Gas
COPL	Canadian Overseas Petroleum Limited	Oil & Gas
LKOH	PJSC Lukoil	Oil & Gas
CNE	Cairn Energy Plc	Oil & Gas

(continued on next page)

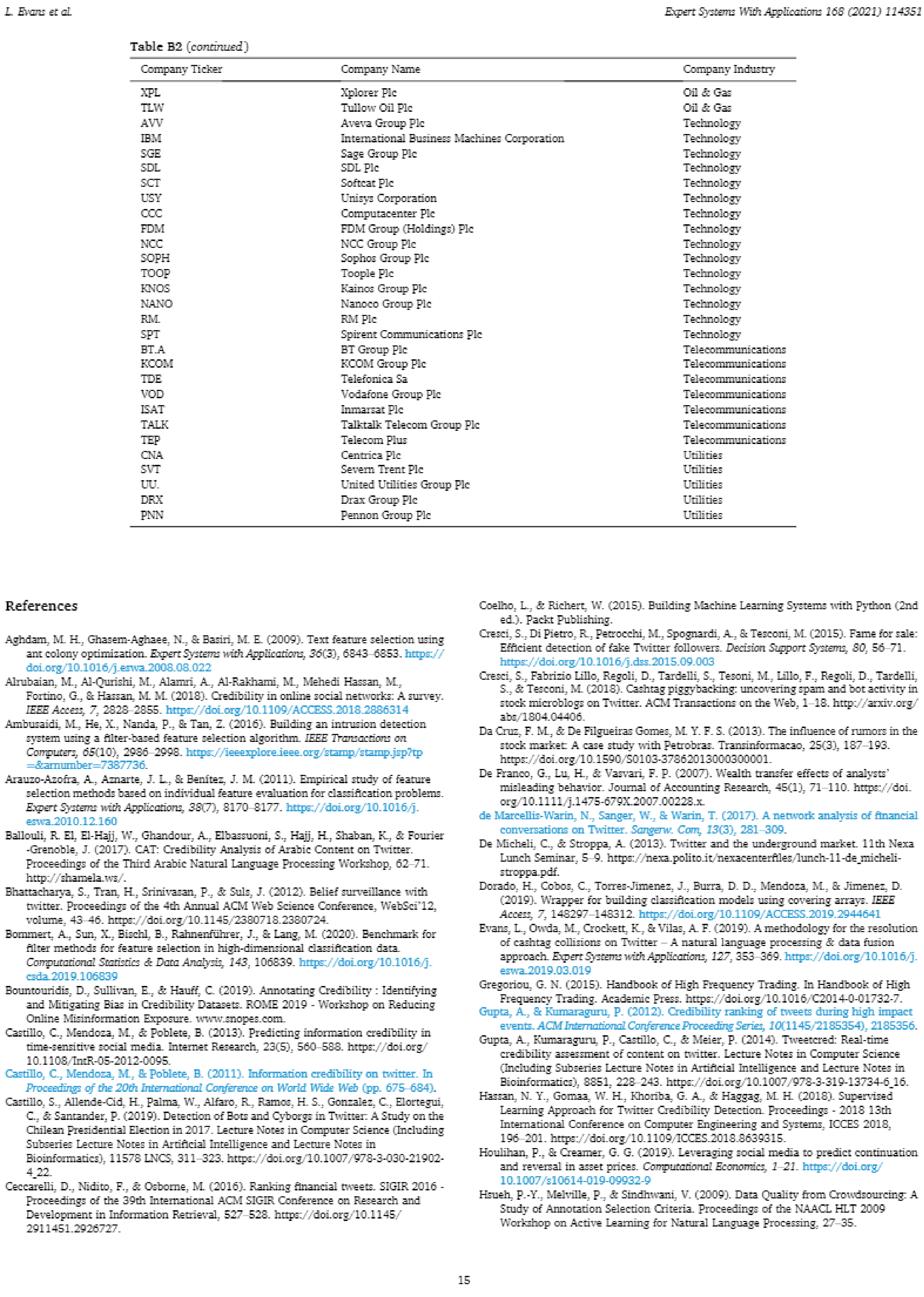


Table B2 (*continued*)

Company Ticker	Company Name	Company Industry
XPL	Xplorer Plc	Oil & Gas
TLW	Tullow Oil Plc	Oil & Gas
AVV	Aveva Group Plc	Technology
IBM	International Business Machines Corporation	Technology
SGE	Sage Group Plc	Technology
SDL	SDL Plc	Technology
SCT	Softcat Plc	Technology
USY	Unisys Corporation	Technology
CCC	Computacenter Plc	Technology
FDM	FDM Group (Holdings) Plc	Technology
NCC	NCC Group Plc	Technology
SOPH	Sophos Group Plc	Technology
TOOP	Toople Plc	Technology
KNOS	Kainos Group Plc	Technology
NANO	Nanoco Group Plc	Technology
RM.	RM Plc	Technology
SPT	Spirent Communications Plc	Technology
BT.A	BT Group Plc	Telecommunications
KCOM	KCOM Group Plc	Telecommunications
TDE	Telefonica Sa	Telecommunications
VOD	Vodafone Group Plc	Telecommunications
ISAT	Inmarsat Plc	Telecommunications
TALK	Talktalk Telecom Group Plc	Telecommunications
TEP	Telecom Plus	Telecommunications
CNA	Centrica Plc	Utilities
SVT	Severn Trent Plc	Utilities
UU.	United Utilities Group Plc	Utilities
DRX	Drax Group Plc	Utilities
PNN	Pennon Group Plc	Utilities

References

Aghdam, M. H., Ghasem-Aghaee, N., & Basiri, M. E. (2009). Text feature selection using ant colony optimization. *Expert Systems with Applications, 36*(3), 6843–6853. https://doi.org/10.1016/j.eswa.2008.08.022

Alrubaian, M., Al-Qurishi, M., Alamri, A., Al-Rakhami, M., Mehedi Hassan, M., Fortino, G., & Hassan, M. M. (2018). Credibility in online social networks: A survey. *IEEE Access, 7*, 2828–2855. https://doi.org/10.1109/ACCESS.2018.2886314

Ambusaidi, M., He, X., Nanda, P., & Tan, Z. (2016). Building an intrusion detection system using a filter-based feature selection algorithm. *IEEE Transactions on Computers, 65*(10), 2986–2998. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7387736.

Arauzo-Azofra, A., Aznarte, J. L., & Benítez, J. M. (2011). Empirical study of feature selection methods based on individual feature evaluation for classification problems. *Expert Systems with Applications, 38*(7), 8170–8177. https://doi.org/10.1016/j.eswa.2010.12.160

Ballouli, R. El, El-Hajj, W., Ghandour, A., Elbassuoni, S., Hajj, H., Shaban, K., & Fourier -Grenoble, J. (2017). CAT: Credibility Analysis of Arabic Content on Twitter. Proceedings of the Third Arabic Natural Language Processing Workshop, 62–71. http://shamela.ws/.

Bhattacharya, S., Tran, H., Srinivasan, P., & Suls, J. (2012). Belief surveillance with twitter. Proceedings of the 4th Annual ACM Web Science Conference, WebSci'12, volume, 43–46. https://doi.org/10.1145/2380718.2380724.

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis, 143*, 106839. https://doi.org/10.1016/j.csda.2019.106839

Bountouridis, D., Sullivan, E., & Hauff, C. (2019). Annotating Credibility : Identifying and Mitigating Bias in Credibility Datasets. ROME 2019 - Workshop on Reducing Online Misinformation Exposure. www.snopes.com.

Castillo, C., Mendoza, M., & Poblete, B. (2013). Predicting information credibility in time-sensitive social media. Internet Research, 23(5), 560–588. https://doi.org/10.1108/IntR-05-2012-0095.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 675–684).

Castillo, S., Allende-Cid, H., Palma, W., Alfaro, R., Ramos, H. S., Gonzalez, C., Elortegui, C., & Santander, P. (2019). Detection of Bots and Cyborgs in Twitter: A Study on the Chilean Presidential Election in 2017. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11578 LNCS, 311–323. https://doi.org/10.1007/978-3-030-21902-4_22.

Ceccarelli, D., Nidito, F., & Osborne, M. (2016). Ranking financial tweets. SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, 527–528. https://doi.org/10.1145/2911451.2926727.

Coelho, L., & Richert, W. (2015). Building Machine Learning Systems with Python (2nd ed.). Packt Publishing.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems, 80*, 56–71. https://doi.org/10.1016/j.dss.2015.09.003

Cresci, S., Fabrizio Lillo, Regoli, D., Tardelli, S., Tesoni, M., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2018). Cashtag piggybacking: uncovering spam and bot activity in stock microblogs on Twitter. ACM Transactions on the Web, 1–18. http://arxiv.org/abs/1804.04406.

Da Cruz, F. M., & De Filgueiras Gomes, M. Y. F. S. (2013). The influence of rumors in the stock market: A case study with Petrobras. Transinformacao, 25(3), 187–193. https://doi.org/10.1590/S0103-37862013000300001.

De Franco, G., Lu, H., & Vasvari, F. P. (2007). Wealth transfer effects of analysts' misleading behavior. Journal of Accounting Research, 45(1), 71–110. https://doi.org/10.1111/j.1475-679X.2007.00228.x.

de Marcellis-Warin, N., Sanger, W., & Warin, T. (2017). A network analysis of financial conversations on Twitter. *Sangerw. Com, 13*(3), 281–309.

De Micheli, C., & Stroppa, A. (2013). Twitter and the underground market. 11th Nexa Lunch Seminar, 5–9. https://nexa.polito.it/nexacenterfiles/lunch-11-de_micheli-stroppa.pdf.

Dorado, H., Cobos, C., Torres-Jimenez, J., Burra, D. D., Mendoza, M., & Jimenez, D. (2019). Wrapper for building classification models using covering arrays. *IEEE Access, 7*, 148297–148312. https://doi.org/10.1109/ACCESS.2019.2944641

Evans, L., Owda, M., Crockett, K., & Vilas, A. F. (2019). A methodology for the resolution of cashtag collisions on Twitter – A natural language processing & data fusion approach. *Expert Systems with Applications, 127*, 353–369. https://doi.org/10.1016/j.eswa.2019.03.019

Gregoriou, G. N. (2015). Handbook of High Frequency Trading. In Handbook of High Frequency Trading. Academic Press. https://doi.org/10.1016/C2014-0-01732-7.

Gupta, A., & Kumaraguru, P. (2012). Credibility ranking of tweets during high impact events. *ACM International Conference Proceeding Series, 10*(1145/2185354), 2185356.

Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014). Tweetcred: Real-time credibility assessment of content on twitter. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8851, 228–243. https://doi.org/10.1007/978-3-319-13734-6_16.

Hassan, N. Y., Gomaa, W. H., Khoriba, G. A., & Haggag, M. H. (2018). Supervised Learning Approach for Twitter Credibility Detection. Proceedings - 2018 13th International Conference on Computer Engineering and Systems, ICCES 2018, 196–201. https://doi.org/10.1109/ICCES.2018.8639315.

Houlihan, P., & Creamer, G. G. (2019). Leveraging social media to predict continuation and reversal in asset prices. *Computational Economics, 1–21*. https://doi.org/10.1007/s10614-019-09932-9

Hsueh, P.-Y., Melville, P., & Sindhwani, V. (2009). Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria. Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, 27–35.

Krzysztof, L., Jacek, S.-W., Michal, J.-L., & Amit, G. (2015). Automated Credibility Assessment on Twitter. Computer Science, 16(2), 157. https://doi.org/10.7494/csci.2015.16.2.157.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159. https://doi.org/10.2307/2529310

Liu, Z., Liu, L.u., & Li, H. (2012). Determinants of information retweeting in microblogging. *Internet Research, 22*(4), 443–466. https://doi.org/10.1108/10662241211250980

Loughran, T., & Mcdonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research, 54(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123.

Loughran, T., Mcdonald, B., Battalio, R., Easton, P., Fuehrmeyer, J., Gao, P., Harvey, C., Hirschey, N., Marietta-Westberg, J., & Schultz, P. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–64. https://www.uts.edu.au/sites/default/files/ADG_Cons2015_Loughran McDonald JE 2011.pdf.

Maddock, J., Starbird, K., Al-Hassani, H., Sandoval, D. E., Orand, M., & Mason, R. M. (2015). Characterizing online rumoring behavior using multi-dimensional signatures. CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing, 228–241. https://doi.org/10.1145/2675133.2675280.

Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012). Tweeting is believing? Understanding microblog credibility perceptions. Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, 441–450. https://doi.org/10.1145/2145204.2145274.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications, 42*(24), 9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052

Odonovan, J., Kang, B., Meyer, G., Hollerer, T., & Adalii, S. (2012). Credibility in context: An analysis of feature distributions in twitter. Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012, 293–301. https://doi.org/10.1109/SocialCom-PASSAT.2012.128.

Oliveira, N., Cortez, P., & Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems, 85*, 62–73. https://doi.org/10.1016/j.dss.2016.02.013

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications, 73*, 125–144. https://doi.org/10.1016/j.eswa.2016.12.036

Page, J. T., & Duffy, M. E. (2018). What Does Credibility Look like? Tweets and Walls in U.S. Presidential Candidates' Visual Storytelling. *Journal of Political Marketing, 17*(1), 3–31. https://doi.org/10.1080/15377857.2016.1171819

Parmezan, A. R. S., Lee, H. D., & Wu, F. C. (2017). Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework. *Expert Systems with Applications, 75*, 1–24. https://doi.org/10.1016/j.eswa.2017.01.013

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of twitter sentiment on stock price returns. PLoS ONE, 10(9), 1–21. https://doi.org/10.1371/journal.pone.0138441.

Rani, D. S., Rani, T. S., & Durga Bhavani, S. (2015). Feature subset selection using consensus clustering. In *ICAPR 2015–2015 8th International Conference on Advances in Pattern Recognition*. https://doi.org/10.1109/ICAPR.2015.7050659

Reidsma, D., & op den Akker, R. (2008). Exploiting "Subjective" Annotations. Workshop on Human Judgements in Computational Linguistics, 8–16.

Ronaghan, S. (2018). The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark. https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3.

Rong, Miao, Gong, Dunwei, & Gao, Xiaozhi (2019). Feature selection and its use in big data: Challenges, methods, and trends. *IEEE Access, 7*, 19709–19725. https://doi.org/10.1109/ACCESS.2019.2894366

Saguna, Zaslavsky, A., & Paris, C. (2012). Context-aware twitter validator (CATVal): A system to validate credibility and authenticity of twitter content for use in decision support systems. Frontiers in Artificial Intelligence and Applications, 238, 323–334. https://doi.org/10.3233/978-1-61499-073-4-323.

Sikdar, S., Adali, S., Amin, M., Abdelzaher, T., Chan, K., Cho, J. H., … O'Donovan, J. (2014). Finding true and credible information on Twitter. In *FUSION 2014–17th International Conference on Information Fusion* (pp. 1–8).

Sikdar, Sujoy, Kang, B., O'donovan, J., Höllerer, T., & Adal, S. (2013). Understanding Information Credibility on Twitter. 2013 International Conference on Social Computing, 19–24. http://www.cs.rpi.edu/~sikdas/papers/socialcom2013.pdf.

Sikdar, Sujoy, Kang, B., O'donovan, J., & Höllerer, T. H. (2013). Cutting Through the Noise: Defining Ground Truth in Information Credibility on Twitter. Human, 2(3), 151–167. https://www.researchgate.net/publication/257200399.

Stringhini, G., Wang, G., Egele, M., Kruegel, C., Vigna, G., Zheng, H., & Zhao, B. Y. (2013). Follow the green: Growth and dynamics in Twitter follower markets. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference* (pp. 163–176). https://doi.org/10.1145/2504730.2504731

Thomson, R., Ito, N., Suda, H., Lin, F., Liu, Y., Hayasaka, R., … Wang, Z. (2012). Trusting tweets: The Fukushima disaster and information source credibility on Twitter. In *ISCRAM 2012 Conference Proceedings - 9th International Conference on Information Systems for Crisis Response and Management* (pp. 1–10).

Tsai, Chih-Fong, & Chen, Yu-Chi (2019). The optimal combination of feature selection and data discretization: An empirical study. *Information Sciences, 505*, 282–293. https://doi.org/10.1016/j.ins.2019.07.091

Yang, F., Liu, Y., Yu, X., & Yang, M. (2012). Automatic Detection of Rumor on Sina Weibo Categories and Subject Descriptors. *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, 2*.

Yang, J., Yu, M., Qin, H., Lu, M., & Yang, C. (2019). A Twitter data credibility framework - hurricane harvey as a use case. *ISPRS International Journal of Geo-Information, 8*(3), 1–21. https://doi.org/10.3390/ijgi8030111

Yang, Min-Chul, & Rim, Hae-Chang (2014). Identifying interesting Twitter contents using topical analysis. *Expert Systems with Applications, 41*(9), 4330–4336. https://doi.org/10.1016/j.eswa.2013.12.051

Yang, Zhao, Sun, Xuezheng, & Hardin, James W. (2011). Testing marginal homogeneity in clustered matched-pair data. *Journal of Statistical Planning and Inference, 141*(3), 1313–1318. https://doi.org/10.1016/j.jspi.2010.10.002

Appendix E

Cashtag Collision Experiment Companies

TABLE E.1: Alternative Investment Market companies (with known collisions)

Company Ticker	Company Name	Sector	Tweets Collected	London South East Posts Collected
88E	88 Energy Limited	Oil Gas	0	51,693
ABC	Abcam PLC	Health Care	1221	9
ARL	Atlantis Resources Limited	Oil Gas	69	194
ASC	ASOS PLC	Consumer Servies	229	58
AVN	Avanti Communications Group PLC	Telecommunications	10	1,871
BKY	Berkeley Energia Limited	Basic Materials	75	1,989
CAKE	Patisserie Holdings PLC	Consumer Services	574	60
COG	Cambridge Cognition Holdings PLC	Health Care	722	14
EMAN	Everyman Media Group PLC	Consumer Services	104	7
EYE	Eagle Eye Solutions Group PLC	Technology	207	7
FLOW	Flowgroup PLC	Industrials	344	8,857
GBP	Global Petroleum Limited	Oil Gas	915	2,969
GGP	Greatland Gold PLC	Basic Materials	400	60,023
GOOD	Good Energy Group PLC	Utilities	1034	4
HRN	Hornby PLC	Consumer Goods	1	17
HUNT	Hunters Property PLC	Financials	7	2
ING	Ingenta PLC	Technology	810	0
INSE	Inspired Energy PLC	Industrials	129	194
MTR	Metal Tiger PLC	Financials	112	6,747
MUL	Mulberry Group PLC	Consumer Goods	3	0
NAK	Nakama Group PLC	Industrials	308	8
PLUS	Plus500 Ltd	Financials	256	216
TRB	Tribal Group PLC	Technology	8	3
VRS	Versarien PLC	Basic Materials	941	4,642
WYN	Wynnstay Group PLC	Consumer Goods	597	2

TABLE E.2: Alternative Investment Market companies (no known collisions)

Company Ticker	Company Name	Sector	Tweets Collected	London South East Posts Collected
BGO	Bango PLC	Technology	3	593
BIOM	Biome Technologies PLC	Basic Materials	1	86
BLV	Belvoir Lettings PLC	Financials	4	5
BOO	Boohoo.Com PLC	Consumer Services	39	7012
CLIN	Clinigen Group PLC	Health Care	534	160
CLON	Clontarf Energy PLC	Oil Gas	58	1532
CRPR	Cropper (James) PLC	Basic Materials	1	9
DX.	Dx (Group) PLC	Industrials	0	732
FEVR	Fevertree Drinks PLC	Consumer Goods	9	729
HZD	Horizon Discovery Group PLC	Health Care	31	16
IMTK	Imaginatik PLC	Technology	2	64
ITQ	Interquest Group PLC	Industrials		28
KOOV	Koovs PLC	Consumer Services	7	1065
LCG	London Capital Group Holdings PLC	Financials	0	442
LWRF	Lightwaverf PLC	Consumer Goods	4	433
MANX	Manx Telecom PLC	Telecommunications	6	9
MYT	Mytrah Energy Limited	Utilities	4	159
NAUT	Nautilus Marine Services PLC	Oil Gas	74	9
PREM	Premier African Minerals Limited	Basic Materials	29	57895
SOU	Sound Energy PLC	Oil Gas	26	40872
TUNE	Focusrite PLC	Consumer Goods	13	10
TUNG	Tungsten Corporation PLC	Financials	10	88
WAND	Wandisco PLC	Technology	691	276
WYG	WYG PLC	Industrials	4	73
YOU	Yougov PLC	Consumer Services	12	2

TABLE E.3: Main Market companies (no known collisions)

Company Ticker	Company Name	Sector	Tweets Collected	London South East Posts Collected
AVV	Aveva Group PLC	Technology	11	5
BARC	Barclays PLC	Financials	822	1738
BBYB	Balfour Beatty PLC	Industrials	0	0
BFA	BASF SE	Basic Materials	11	0
BP.	BP PLC	Oil Gas	0	833
BT.A	BT Group PLC	Telecommunications	52	7660
DEB	Debenhams PLC	Consumer Services	755	1109
ECM	Electrocomponents PLC	Industrials	20	3
GNS	Genus PLC	Health Care	7	4
HFD	Halfords Group PLC	Consumer Services	8	62
HSBA	HSBC Holdings PLC	Financials	170	386
KCOM	KCOM Group PLC	Telecommunications	7	46
MRW	Morrison (Wm) Supermarkets PLC	Consumer Services	57	120
OXB	Oxford Biomedica PLC	Health Care	29	914
PDL	Petra Diamonds Limited	Basic Materials	58	568
PSN	Persimmon PLC	Consumer Goods	28	43
RR.	Rolls-Royce Holdings PLC	Industrials	0	375
SGE	Sage Group PLC	Technology	44	17
SHP	Shire PLC	Health Care	1048	759
TYT	Toyota Motor Corporation	Consumer Goods	2	0
UAI	U and I Group PLC	Financials	7	38
USY	Unisys Corporation	Technology	1	0
UU.	United Utilities Group PLC	Utilities	0	101
WG.	Wood Group (John) PLC	Oil Gas	0	70
ZCC	ZCCM Investments Holdings PLC	Basic Materials	57	0

Appendix F

Credibility Classifier Features

TABLE F.1: General feature list

Feature Sub-Group	Feature Num.	Feature	Notes
Content	1	Tweet Length (Chars)	Length of the tweet in characters (including spaces)
	2	Tweet Length (Words)	Length of the tweet in words
	3	Tweet Contains Question Mark (QM)	Does the tweet contain a question mark
	4	Tweet Contains Multiple QMs	Does the tweet contain multiple question marks
	5	Tweet Contains Exclamation Mark (EM)	Does the tweet contain an exclamation mark
	6	Tweet Contains Multiple EMs	Does the tweet contain multiple exclamation marks
	7	Tweet Contains First Person Pronouns	e.g. I, we, us, me, my, mine, our, ours
	8	Tweet Contains Second Person Pronouns	e.g. you, your, yours
	9	Tweet Contains Third Person Pronouns	e.g. he, she, her, him, it, they, them, theirs
	10	Tweet Contains Positive Emoticons	e.g. :), :-)
	11	Tweet Contains Negative Emoticons	e.g. :(, :-(
	12	Tweet Contains User Mention	Does the tweet contain an @ user mention
	13	Tweet Hashtag Count	The count of word prefixed with a hashtag (#) as determined by the tweet JSON object
	14	Is Retweet (RT)	Contains RT at the start of the tweet text

	15	URL Count	The count of URLs within the tweet
	16	Per cent Uppercase	The percentage of the tweet which is in UPPERCASE
	17	Is Quote Tweet	If the tweet is quoting (e.g. replying) to another tweet
	18	Contains Media	Contains an image, video or gif
	19	Present Verb Count	Count of verbs in present tense within the tweet text
	20	Past Verb Count	Count of verbs in past tense within the tweet text
	21	Adjective Count	Count of adjectives within the tweet text
	22	Interjection Count	Count of interjections within the tweet text
	23	Noun Count	Count of nouns within the tweet text
	24	Adverb Count	Count of adverbs within the tweet text
	25	Proper Noun Count	Count of proper nouns within the tweet text
	26	Numerical Cardinal Count	Count of numerical cardinal values within the tweet text
Context	27	Live URL Count	The count of URLs in the tweet which resulted in a successful web response (200)
	28	Tweeted on Weekday	If the tweet was tweeted on a weekday
	29	Top 500 URL Count	As defined by https://moz.com/top500
	30	Tweet Source	0 – Official Twitter Web Client 1 – Twitter for Android 2 – Twitter for iPhone 3 – Automated Tool (e.g. Zapier, IFTTT, Hootsuite, TweetDeck) 4 – Other
User	31	User Account Age (at time of tweet)	The number of days an account has been active on the Twitter platform from when the tweet was published to Twitter
	32	User has URL on Profile	Does the user have a URL on their profile?
	33	User has Default Profile Pic	Is the user using the default profile image provided by Twitter upon registering their account
	34	User has set a Location	Has the user set a location on their profile?
	35	User Verified	Is the user a verified user (blue tick verification seal)?
	36	User Num of Tweets	The number of tweets the user has made (at the time the tweet was collected)
	37	User Follower Count	The number of followers the user's account has
	38	User Following Count	The number of accounts the user is following
	39	User Listed Count	How many lists is the user account's listed on?

	40	User has Desc	Does the user have a description on their profile page?
	41	User Description Length	The length of the user description, 0 if none
	42	User has Real Location	Does the user have a factual location?
	43	Username Length	Length of the user's username
	44	Username Words	The number of words comprising the user name

TABLE F.2: Financial feature list

Feature Sub-Group	Feature Num.	Feature	Notes
Content	45	Count of positive financial keywords	As defined by research by Loughran and McDonald (2011).
	46	Count of negative financial keywords	
	47	Count of uncertainty financial keywords	
	48	Count of litigious financial keywords	
	49	Count of constraining financial keywords	
Company-Specific Features	50	Close – Open Price (range) on day	Provided by the AlphaVantage API
	51	High – Low Price (range) on day	
	52	RNS published on day	Was a Regulatory News Service (RNS) statement issued for the company corresponding to the first experiment cashtag encountered on the day the tweet was made?
	53	Broker Rating issued on day	Was a Broker rating issued for the company corresponding to the first experiment cashtag encountered on the day the tweet was made?

Exchange-Specific Features	54	Credible Fin URLs in Tweet	A list of URLs found to be credible investment or news websites, hand-curated by an expert based on all the URLs found occurring in at least 1% of the overall tweets collected.
	55	Tweeted Before Market Open	These features differ depending on the stock exchange.
	56	Tweeted During Market Open	
	57	Tweeted After Market Closed	
	58	Count Cashtags (CTs)	
	59+	Count of each industry Cashtags	

Appendix G

Detection Layer Clustering Features

G.1 Event Clustering Features

TABLE G.1: Event clustering features

Feature	**Description**
Pre-event total tweets	The number of LSE tweets in the pre-event window.
Pre-event total credible tweets	The number of LSE tweets in the pre-event window that were classified as being credible
Pre-event total ambiguous tweets	The number of LSE tweets in the pre-event window that were classified as being ambiguous
Pre-event total not credible tweets	The number of LSE tweets in the pre-event window that were classified as being not credible
Pre-event total FDB posts	The total number of FDB posts in the pre-event window
Pre-event total unique Twitter users	The number of unique Twitter users participating in discussion during the pre-event window
Pre-event total unique FDB users	The number of unique FDB users participating in discussion during the pre-event window
Post-event total tweets	The number of LSE tweets in the post-event window
Post-event total credible tweets	The number of LSE tweets in the post-event window that were classified as being credible
Post-event total ambiguous tweets	The number of LSE tweets in the post-event window that were classified as being ambiguous
Post-event total not credible tweets	The number of LSE tweets in the post-event window that were classified as being not credible
Post-event total FDB posts	The total number of FDB posts in the post-event window
Post-event total unique Twitter users	The number of unique Twitter users participating in discussion during the post-event window
Post-event total unique FDB users	The number of unique FDB users participating in discussion during the post-event window

TABLE G.2: Financial Discussion Board (FDB) post clustering features

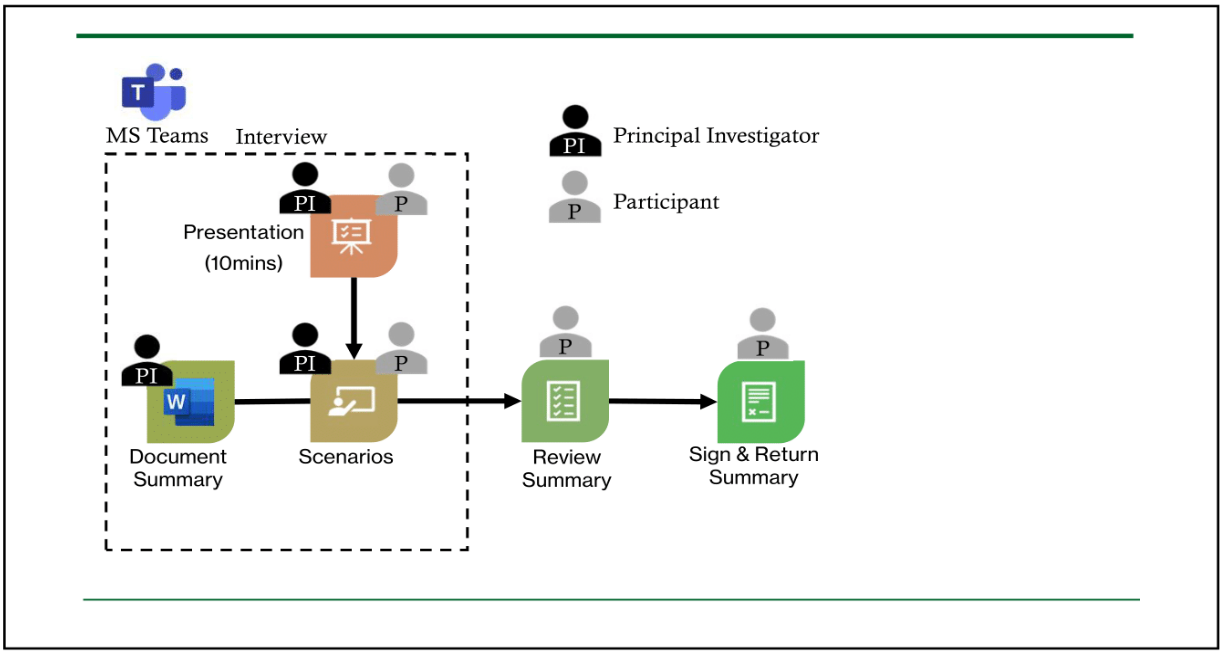
Feature	Description
Count of irregular keywords	The number of keywords that could be considered irregular. This keyword list is based on previous research and
Premium member status	Is the user who made the FDB post a premium member?
Post length (chars)	The total number of characters within the FDB post
Count of URLs	The total number of hyperlinks present within the FDB post
Recommendation count	The total number of recommendations the post has (upvotes by other users of the FDB)
Opinion	The opinion of the poster - this is a dropdown menu located on the London South East FDB, and includes: No Opinion, Strong Buy, Weak Buy, Buy, Hold, Sell, Weak Sell, and Strong Sell
Stock price (at time of posting)	The stock price at the time the FDB post was made
Posting during trading time	Was the post made during London Stock Exchange trading hours?
Author post count	The total number of posts made be the user who published the FDB post (at the time of that post)

TABLE G.3: Tweet clustering features

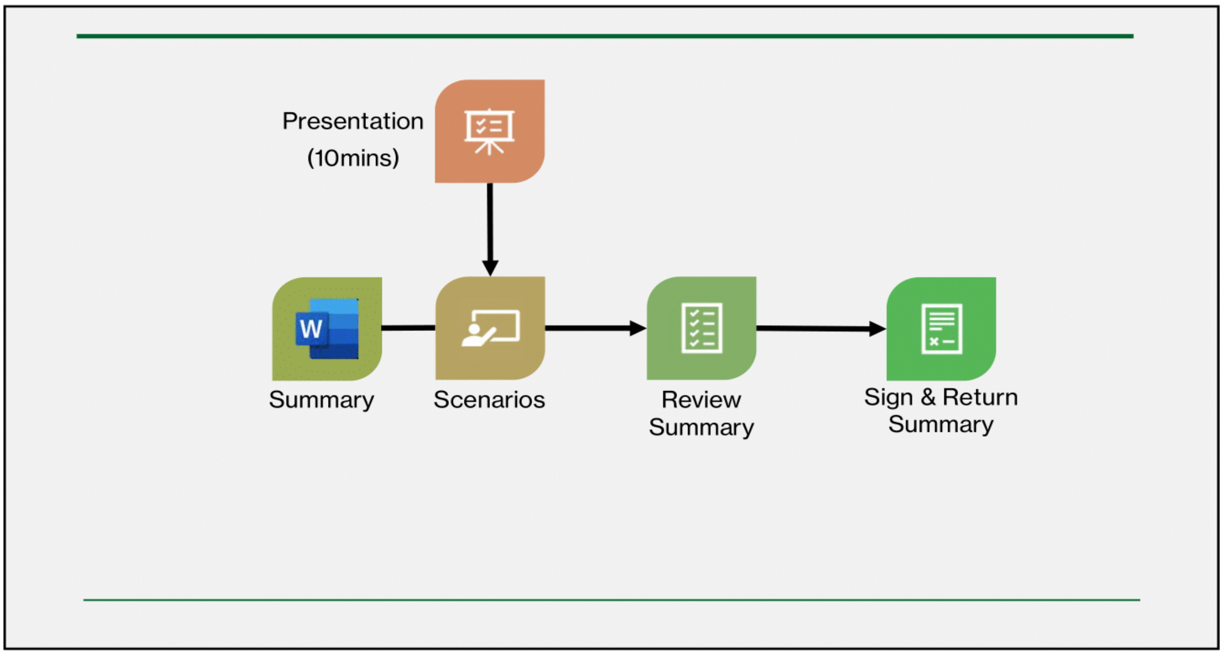
Feature	Description
Count of irregular keywords	The number of keywords that could be considered irregular. This keyword list is based on previous research and
Tweeted during trading	Was the tweet made during London Stock Exchange trading hours?
Cashtag count	The number of cashtags in the tweet
Hashtag count	The number of hashtags in the tweet
Tweet length (chars)	The length of the tweet in characters
Count of URLs	The total number of hyperlinks in the tweet
Tweet Credibility	The credibility of the tweet, as determined by the SDEs credibility classifier.
Media (e.g. image) count	Does the tweet contain some kind of media – e.g. an image, video, or GIF?
Tweet contains user mention	Does the tweet contain a mention of another user?

Appendix H

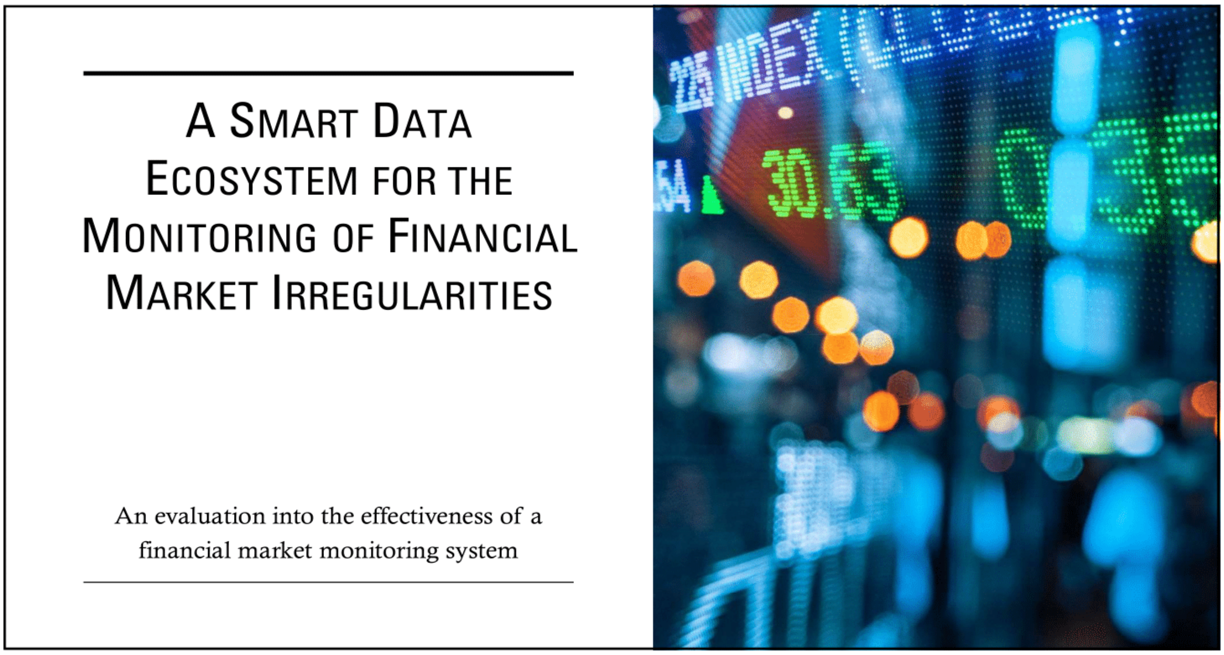
Participant Presentation


MS Teams
Interview
PI
Principal Investigator
P
Participant
Presentation
(10mins)
PI
P
PI
W
Document
Summary
Scenarios
P
Review
Summary
P
Sign & Return
Summary

1


Presentation
(10mins)
W
Summary
Scenarios
Review
Summary
Sign & Return
Summary

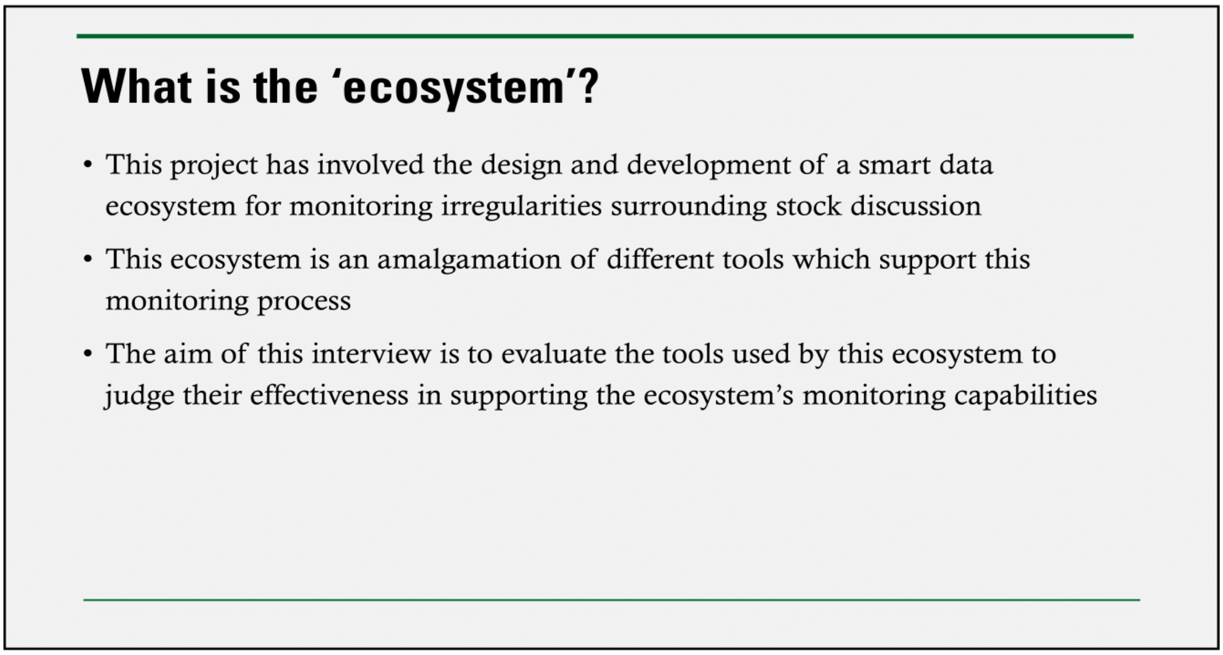
2



A Smart Data Ecosystem for the Monitoring of Financial Market Irregularities

An evaluation into the effectiveness of a financial market monitoring system

3

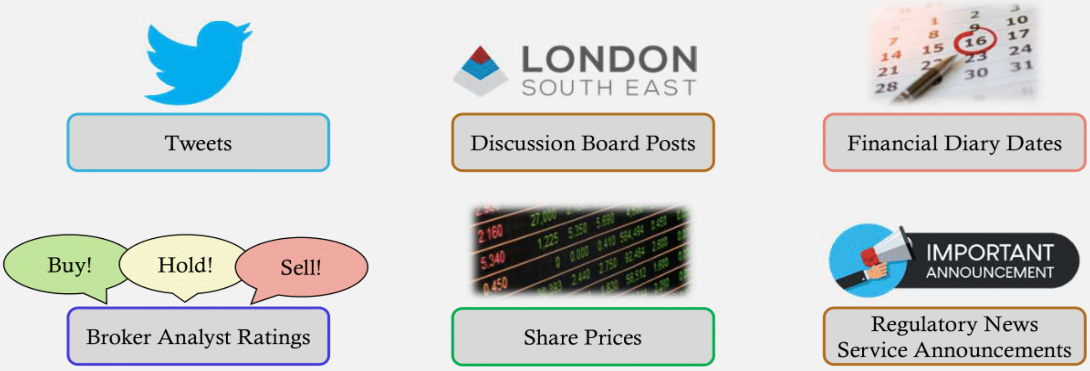


What is the ‘ecosystem’?

- This project has involved the design and development of a smart data ecosystem for monitoring irregularities surrounding stock discussion
- This ecosystem is an amalgamation of different tools which support this monitoring process
- The aim of this interview is to evaluate the tools used by this ecosystem to judge their effectiveness in supporting the ecosystem’s monitoring capabilities

4

Ecosystem Data Sources



Tweets

LONDON SOUTH EAST

Discussion Board Posts

Financial Diary Dates

Buy! Hold! Sell!

Broker Analyst Ratings

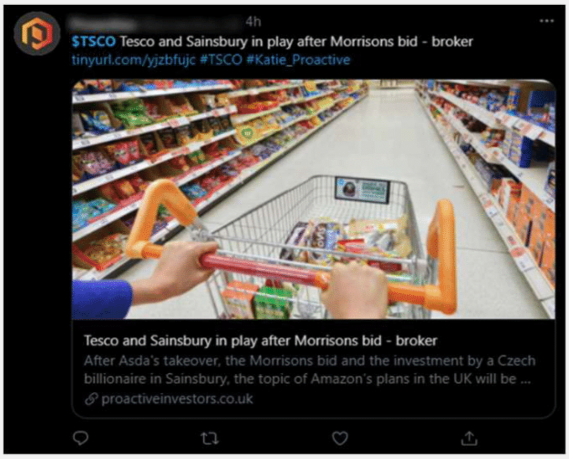
Share Prices

IMPORTANT ANNOUNCEMENT

Regulatory News Service Announcements

5

Example Financial Stock Tweet



[redacted] 4h

$TSCO Tesco and Sainsbury in play after Morrisons bid - broker tinyurl.com/yjzbfujc #TSCO #Katie_Proactive

Tesco and Sainsbury in play after Morrisons bid - broker

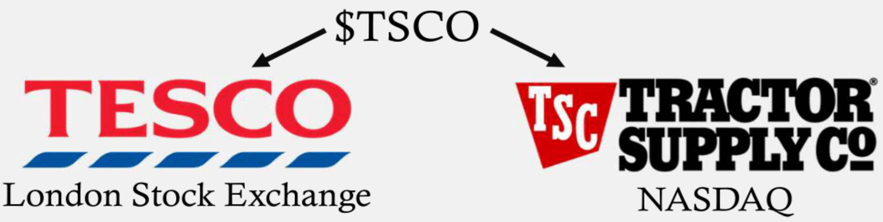
After Asda's takeover, the Morrisons bid and the investment by a Czech billionaire in Sainsbury, the topic of Amazon's plans in the UK will be ...

proactiveinvestors.co.uk

6

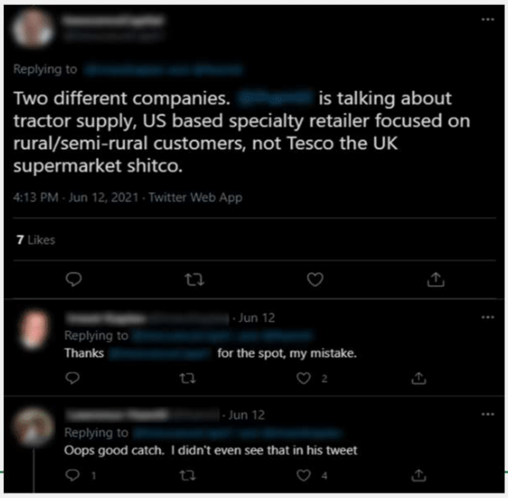
Financial Stock Tweets

- Tweets containing cashtags allow investors to clearly align their tweets with specific stocks – e.g. $TSCO maps with Tesco's TSCO ticker symbol
- Twitter does **not** distinguish between companies on different stock exchanges which have an identical ticker symbol (cashtag)



7

Cashtag Collision Issue for Investors



[redacted]

Replying to [redacted]

Two different companies. [redacted] is talking about tractor supply, US based specialty retailer focused on rural/semi-rural customers, not Tesco the UK supermarket shitco.

4:13 PM · Jun 12, 2021 · Twitter Web App

7 Likes

[redacted] · Jun 12

Replying to [redacted]

Thanks [redacted] for the spot, my mistake.

2

[redacted] · Jun 12

Replying to [redacted]

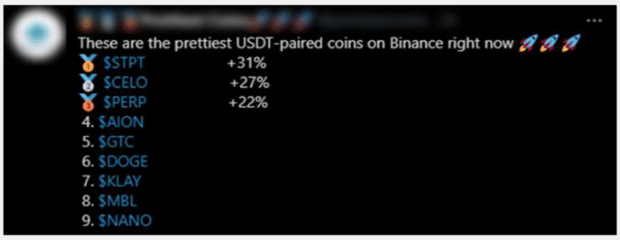
Oops good catch. I didn't even see that in his tweet

1 4

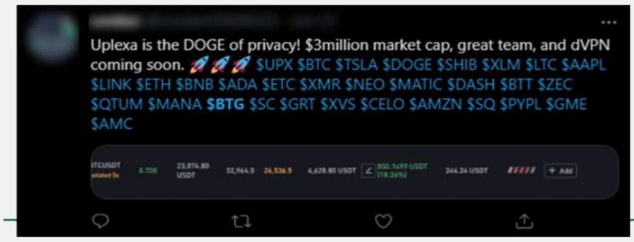
8

19/01/2022

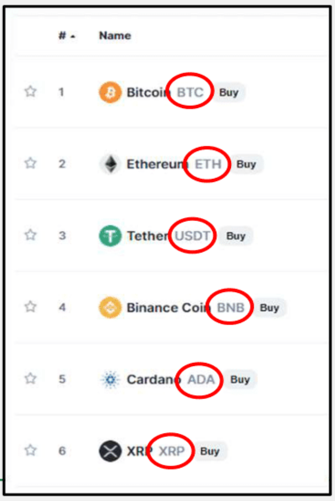
Exacerbating the Issue: Cryptocurrencies



These are the prettiest USDT-paired coins on Binance right now 🚀🚀🚀
🥇 $STPT +31%
🥈 $CELO +27%
🥉 $PERP +22%
4. $AION
5. $GTC
6. $DOGE
7. $KLAY
8. $MBL
9. $NANO



Uplexa is the DOGE of privacy! $3million market cap, great team, and dVPN coming soon. 🚀🚀🚀 $UPX $BTC $TSLA $DOGE $SHIB $XLM $LTC $AAPL $LINK $ETH $BNB $ADA $ETC $XMR $NEO $MATIC $DASH $BTT $ZEC $QTUM $MANA **$BTG** $SC $GRT $XVS $CELO $AMZN $SQ $PYPL $GME $AMC



#	Name
1	Bitcoin BTC Buy
2	Ethereum ETH Buy
3	Tether USDT Buy
4	Binance Coin BNB Buy
5	Cardano ADA Buy
6	XRP XRP Buy

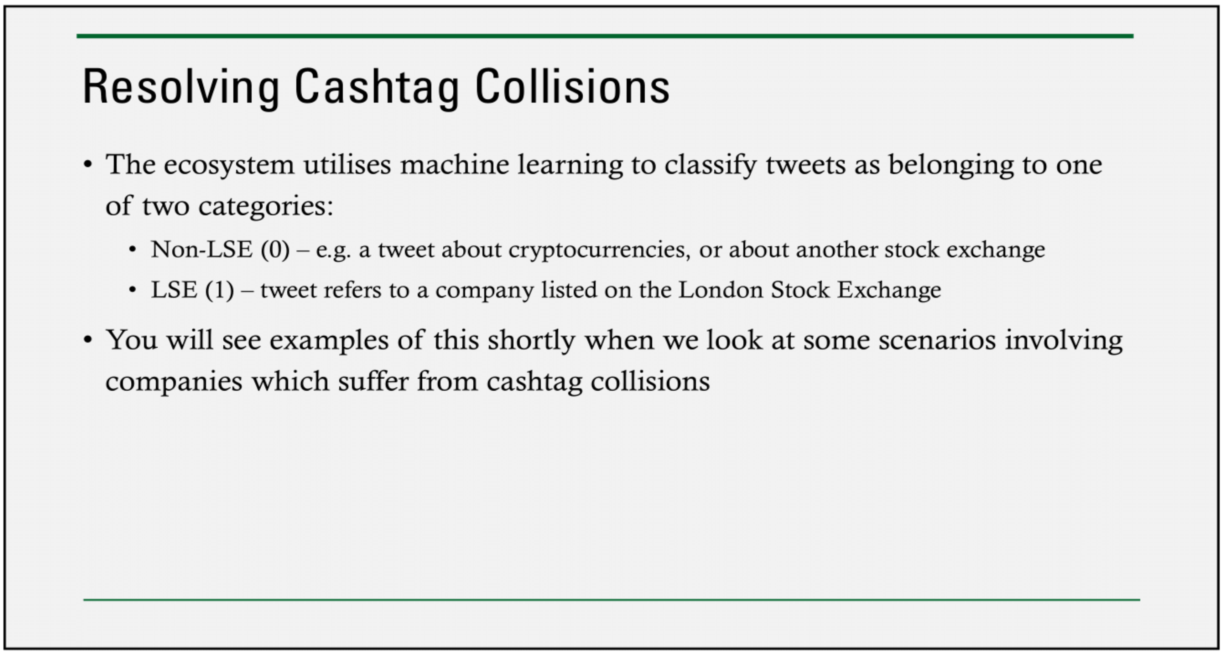
9

How common is this issue?

- Comparing the 2,017 companies listed on the LSE with 7,700+ companies on the NASDAQ:
  - 344 companies (17%) on the LSE share a ticker (and as a result, a cashtag) with companies on the NASDAQ

Ticker	LSE Company	NASDAQ Company
AAL	Anglo American PLC	American Airlines Group Inc.
BLU	Blue Star Capital PLC	BELLUS Health Inc.
CNR	Condor Gold PLCCornerstone	Building Brands Inc.
DFS	DFS Furniture PLC	Discover Financial Services
ESNT	Essentra PLC	Essent Group Ltd.
FLO	Flowtech Fluidpower PLC	Flowers Foods Inc.
GOOD	Good Energy Group PLC	Gladstone Commercial Trust
HUM	Hummingbird Resources PLC	Humana Inc.

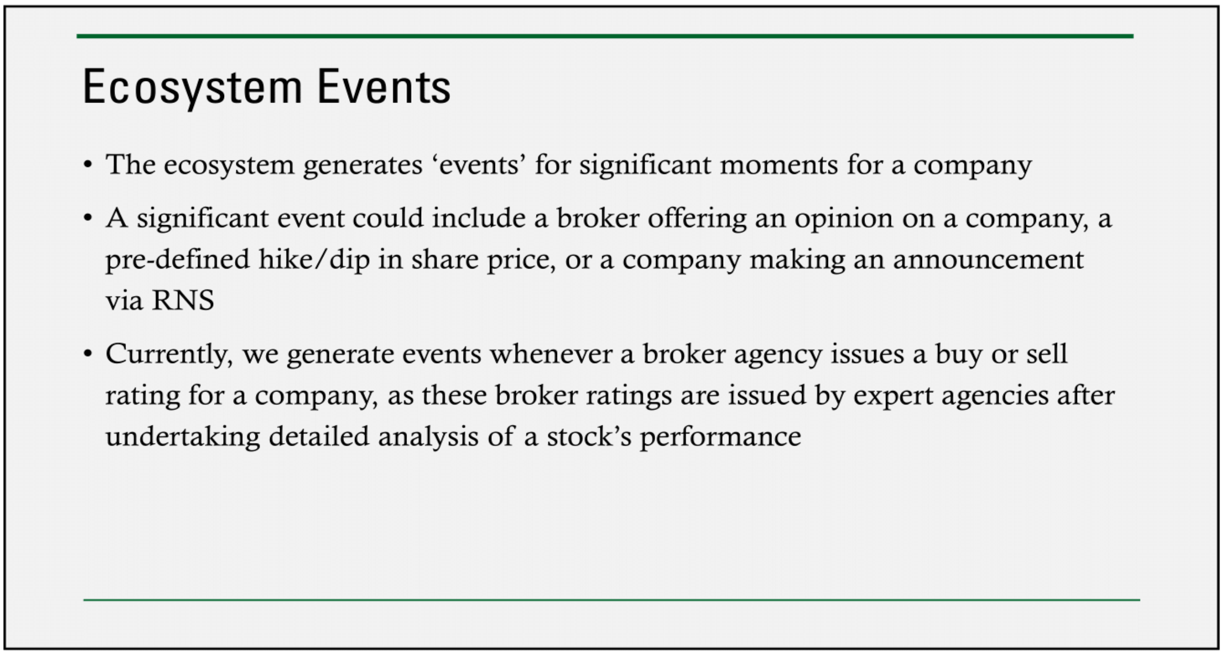
10



Resolving Cashtag Collisions

- The ecosystem utilises machine learning to classify tweets as belonging to one of two categories:
  - Non-LSE (0) – e.g. a tweet about cryptocurrencies, or about another stock exchange
  - LSE (1) – tweet refers to a company listed on the London Stock Exchange
- You will see examples of this shortly when we look at some scenarios involving companies which suffer from cashtag collisions

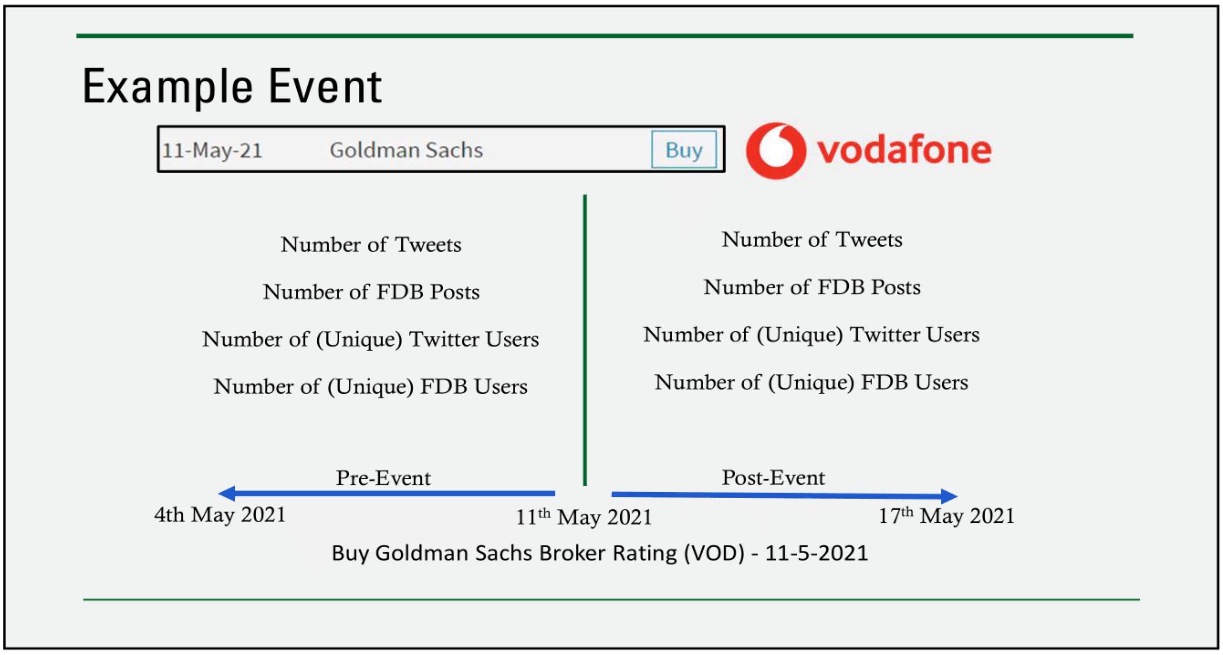
11



Ecosystem Events

- The ecosystem generates ‘events’ for significant moments for a company
- A significant event could include a broker offering an opinion on a company, a pre-defined hike/dip in share price, or a company making an announcement via RNS
- Currently, we generate events whenever a broker agency issues a buy or sell rating for a company, as these broker ratings are issued by expert agencies after undertaking detailed analysis of a stock’s performance

12



Example Event

11-May-21	Goldman Sachs	Buy

Number of Tweets

Number of FDB Posts

Number of (Unique) Twitter Users

Number of (Unique) FDB Users

Number of Tweets

Number of FDB Posts

Number of (Unique) Twitter Users

Number of (Unique) FDB Users

Pre-Event

Post-Event

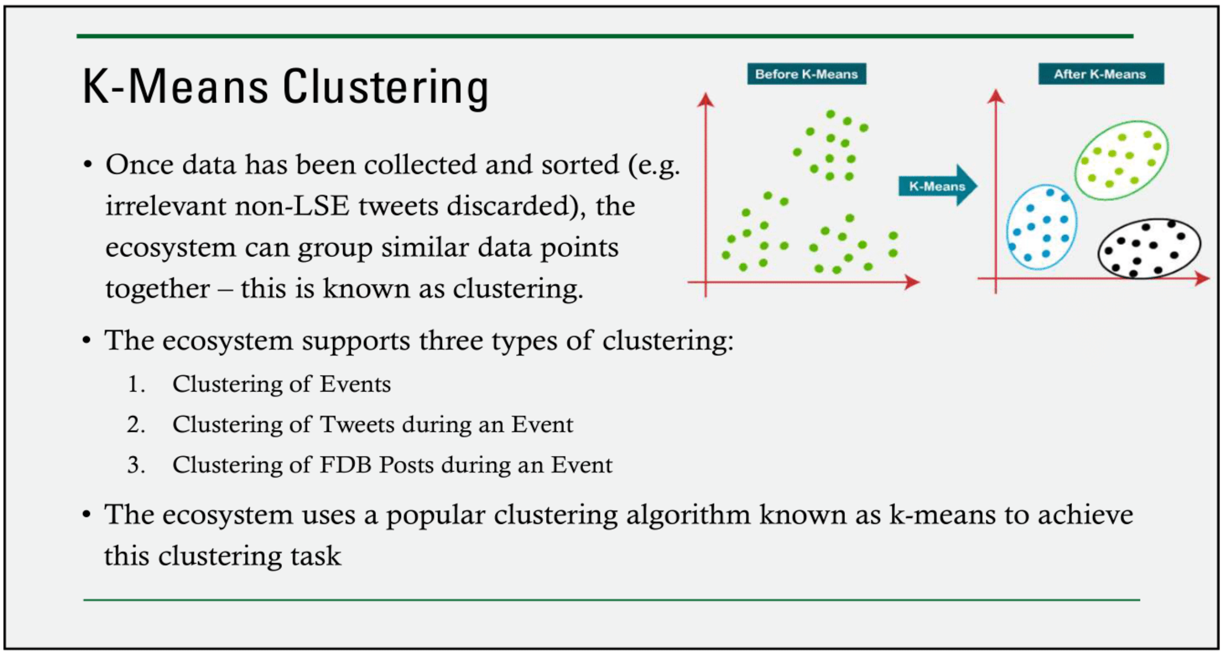
4th May 2021

11th May 2021

17th May 2021

Buy Goldman Sachs Broker Rating (VOD) - 11-5-2021

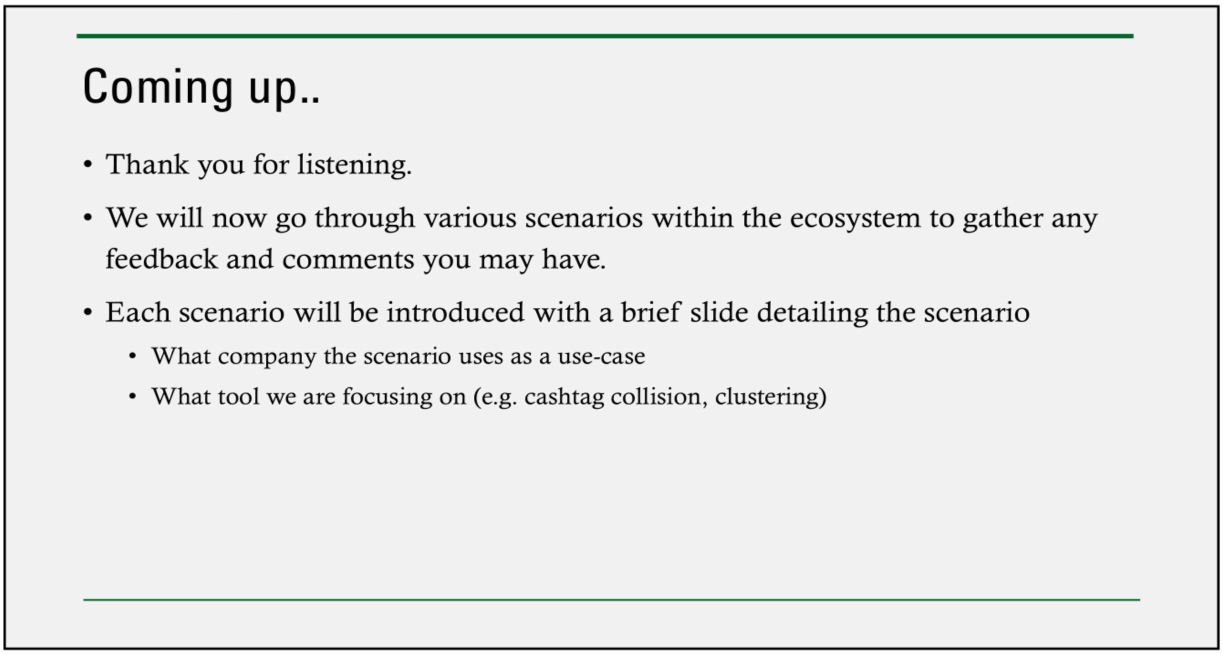
13



K-Means Clustering

- Once data has been collected and sorted (e.g. irrelevant non-LSE tweets discarded), the ecosystem can group similar data points together – this is known as clustering.
- The ecosystem supports three types of clustering:
  1. Clustering of Events
  2. Clustering of Tweets during an Event
  3. Clustering of FDB Posts during an Event
- The ecosystem uses a popular clustering algorithm known as k-means to achieve this clustering task

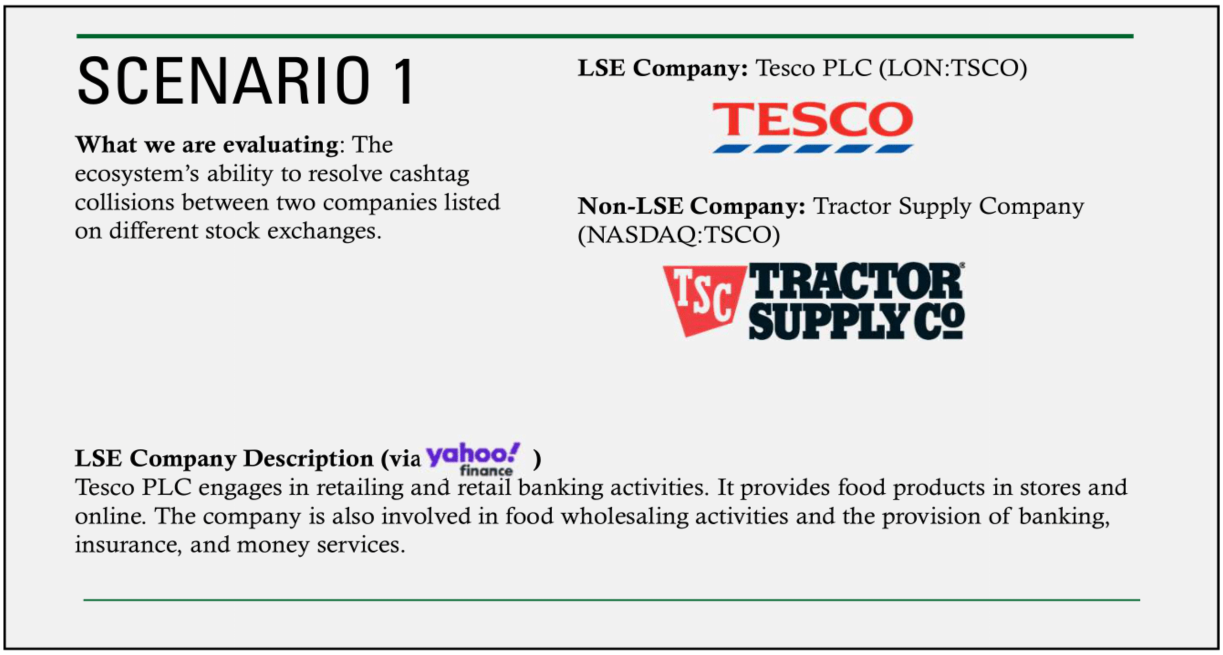
14



Coming up..

- Thank you for listening.
- We will now go through various scenarios within the ecosystem to gather any feedback and comments you may have.
- Each scenario will be introduced with a brief slide detailing the scenario
  - What company the scenario uses as a use-case
  - What tool we are focusing on (e.g. cashtag collision, clustering)

15



SCENARIO 1

What we are evaluating: The ecosystem's ability to resolve cashtag collisions between two companies listed on different stock exchanges.

LSE Company: Tesco PLC (LON:TSCO)

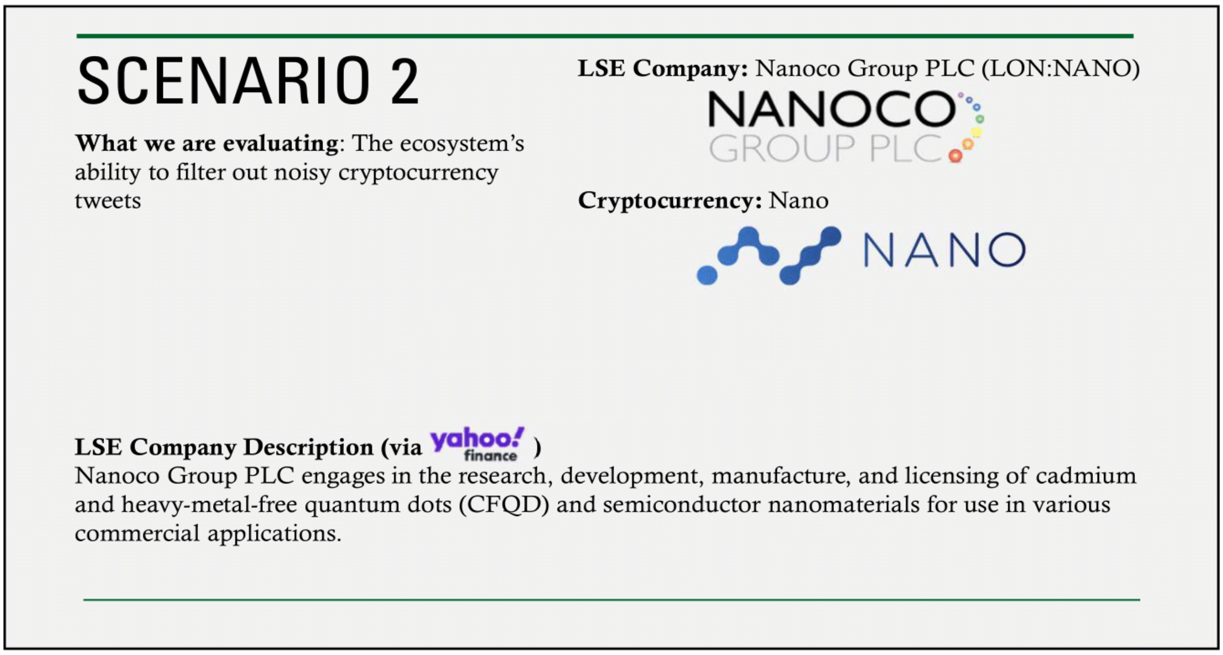
TESCO

Non-LSE Company: Tractor Supply Company (NASDAQ:TSCO)

TSC TRACTOR SUPPLY CO

LSE Company Description (via yahoo! finance)
Tesco PLC engages in retailing and retail banking activities. It provides food products in stores and online. The company is also involved in food wholesaling activities and the provision of banking, insurance, and money services.

16



SCENARIO 2

What we are evaluating: The ecosystem's ability to filter out noisy cryptocurrency tweets

LSE Company: Nanoco Group PLC (LON:NANO)

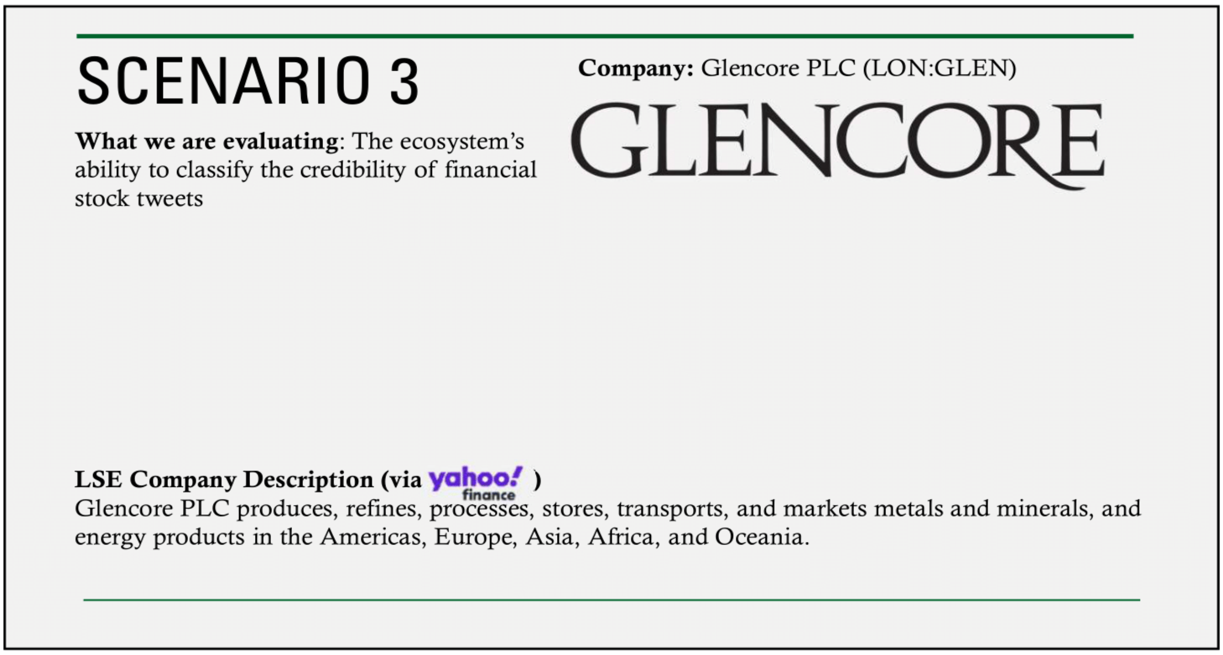
NANOCO GROUP PLC

Cryptocurrency: Nano

NANO

LSE Company Description (via yahoo! finance)
Nanoco Group PLC engages in the research, development, manufacture, and licensing of cadmium and heavy-metal-free quantum dots (CFQD) and semiconductor nanomaterials for use in various commercial applications.

17



SCENARIO 3

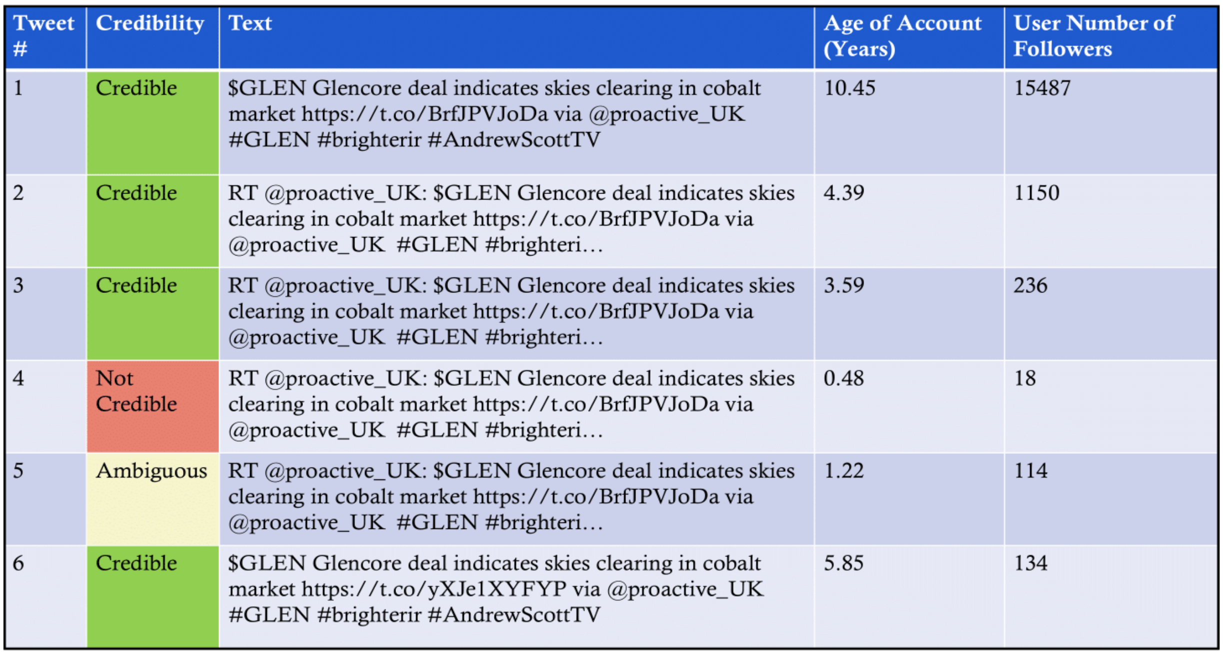
What we are evaluating: The ecosystem's ability to classify the credibility of financial stock tweets

Company: Glencore PLC (LON:GLEN)

GLENCORE

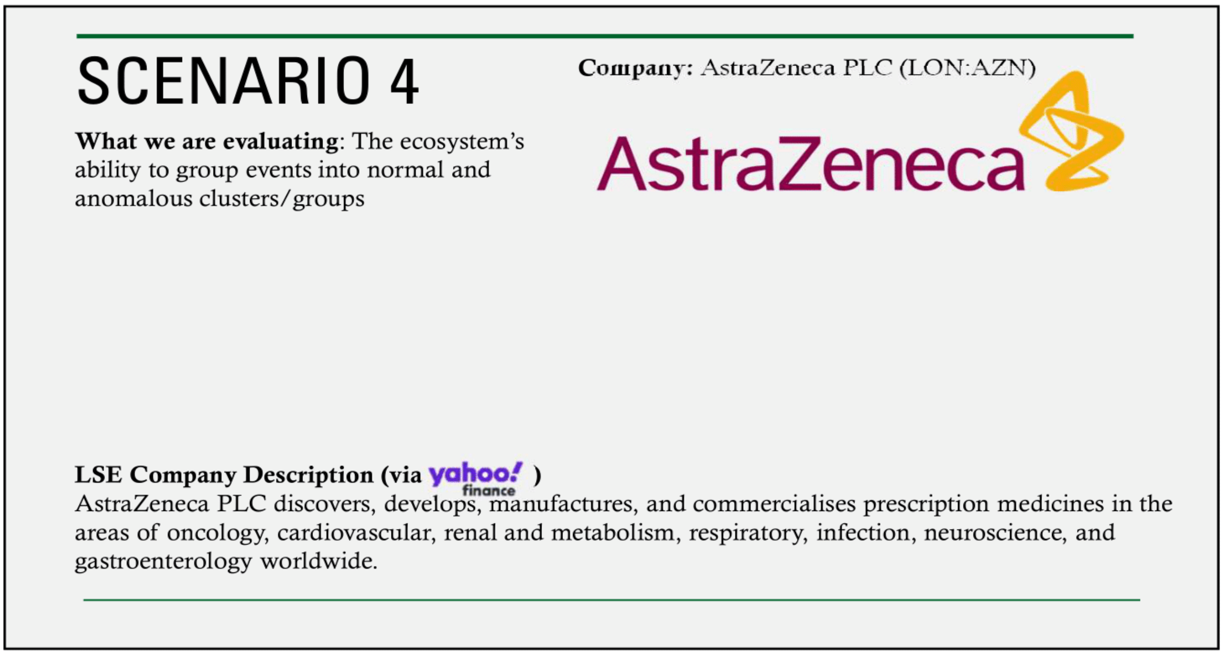
LSE Company Description (via yahoo! finance)
Glencore PLC produces, refines, processes, stores, transports, and markets metals and minerals, and energy products in the Americas, Europe, Asia, Africa, and Oceania.

18



Tweet #	Credibility	Text	Age of Account (Years)	User Number of Followers
1	Credible	$GLEN Glencore deal indicates skies clearing in cobalt market https://t.co/BrfJPVJoDa via @proactive_UK #GLEN #brighterir #AndrewScottTV	10.45	15487
2	Credible	RT @proactive_UK: $GLEN Glencore deal indicates skies clearing in cobalt market https://t.co/BrfJPVJoDa via @proactive_UK #GLEN #brighteri…	4.39	1150
3	Credible	RT @proactive_UK: $GLEN Glencore deal indicates skies clearing in cobalt market https://t.co/BrfJPVJoDa via @proactive_UK #GLEN #brighteri…	3.59	236
4	Not Credible	RT @proactive_UK: $GLEN Glencore deal indicates skies clearing in cobalt market https://t.co/BrfJPVJoDa via @proactive_UK #GLEN #brighteri…	0.48	18
5	Ambiguous	RT @proactive_UK: $GLEN Glencore deal indicates skies clearing in cobalt market https://t.co/BrfJPVJoDa via @proactive_UK #GLEN #brighteri…	1.22	114
6	Credible	$GLEN Glencore deal indicates skies clearing in cobalt market https://t.co/yXJe1XYFYP via @proactive_UK #GLEN #brighterir #AndrewScottTV	5.85	134

19



SCENARIO 4

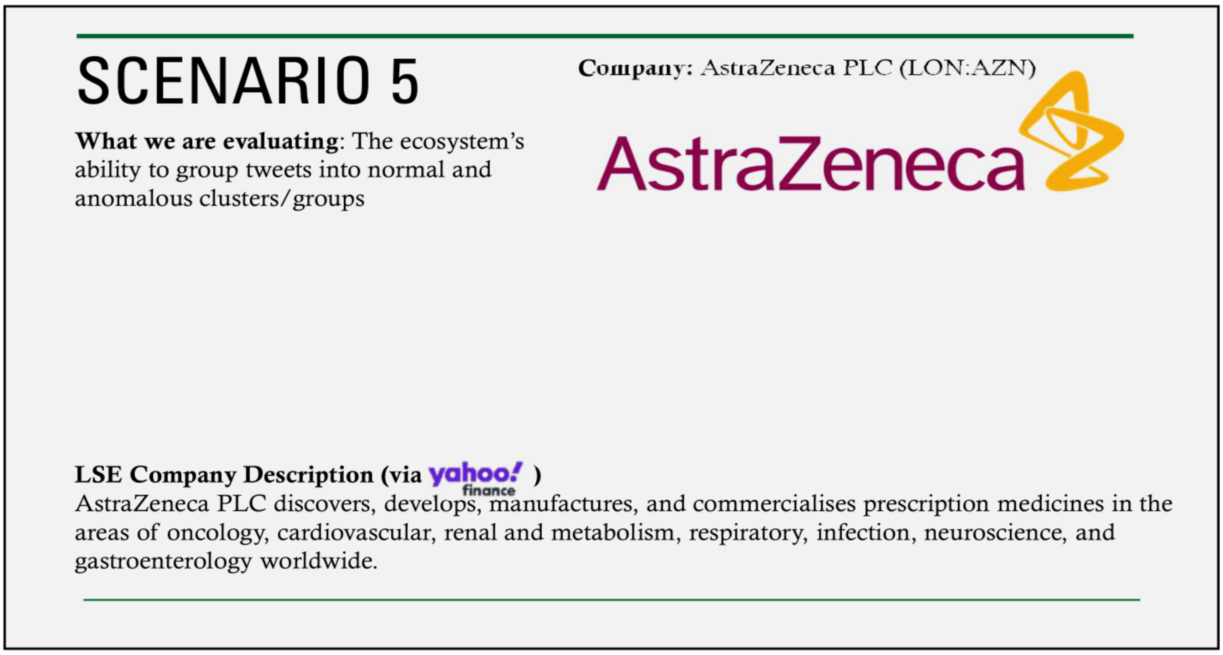
What we are evaluating: The ecosystem's ability to group events into normal and anomalous clusters/groups

Company: AstraZeneca PLC (LON:AZN)

AstraZeneca

LSE Company Description (via yahoo! finance)
AstraZeneca PLC discovers, develops, manufactures, and commercialises prescription medicines in the areas of oncology, cardiovascular, renal and metabolism, respiratory, infection, neuroscience, and gastroenterology worldwide.

20



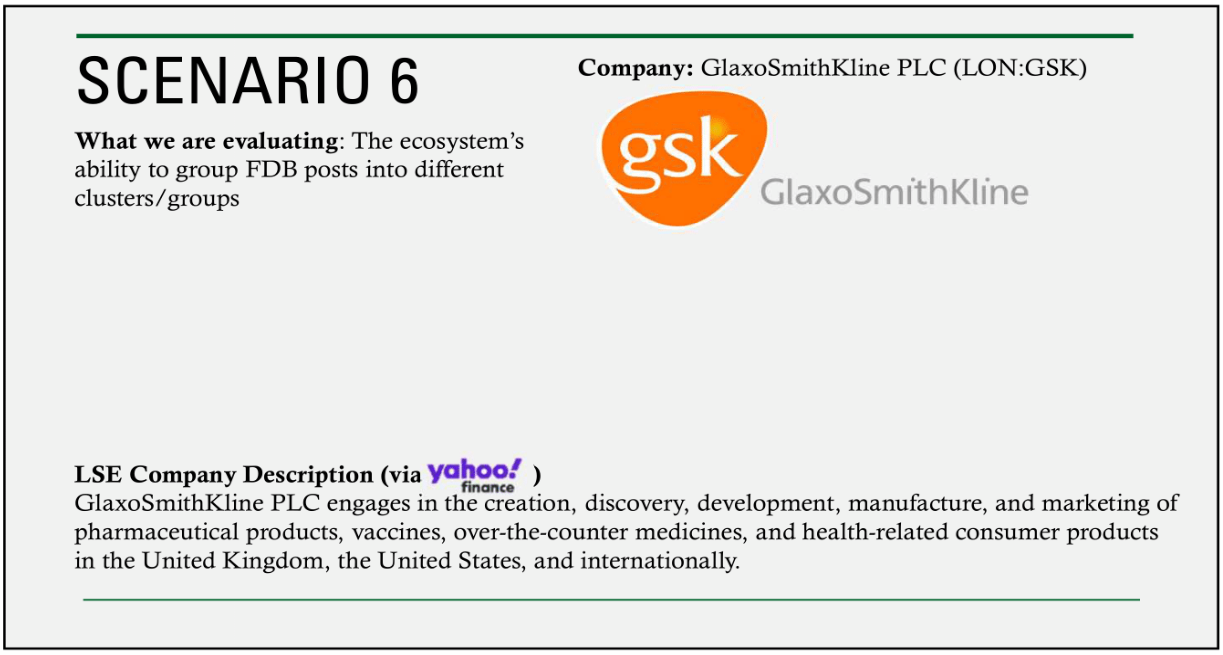
SCENARIO 5

Company: AstraZeneca PLC (LON:AZN)

What we are evaluating: The ecosystem's ability to group tweets into normal and anomalous clusters/groups

LSE Company Description (via yahoo! finance)
AstraZeneca PLC discovers, develops, manufactures, and commercialises prescription medicines in the areas of oncology, cardiovascular, renal and metabolism, respiratory, infection, neuroscience, and gastroenterology worldwide.

21



SCENARIO 6

Company: GlaxoSmithKline PLC (LON:GSK)

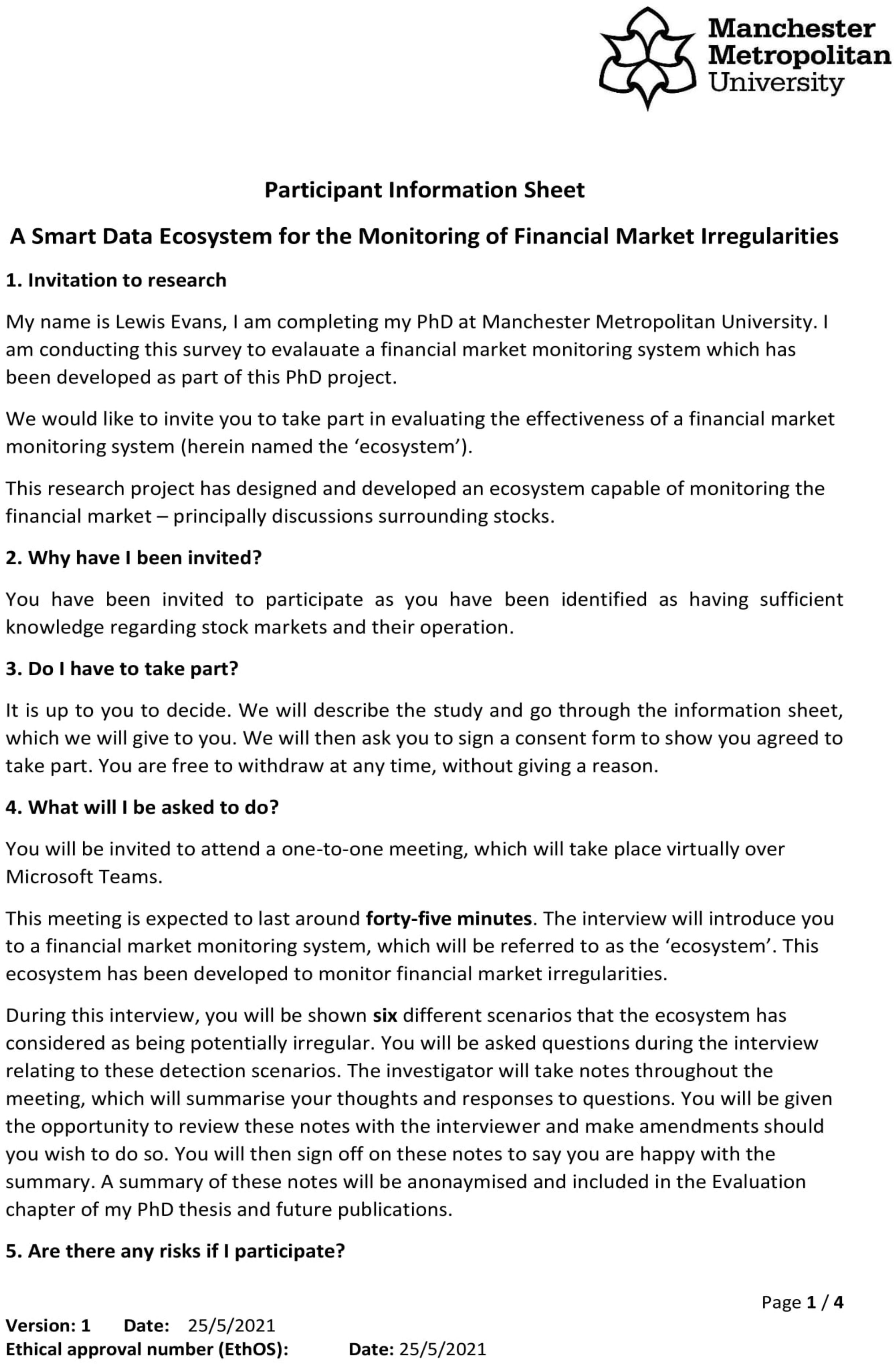
What we are evaluating: The ecosystem's ability to group FDB posts into different clusters/groups

LSE Company Description (via yahoo! finance)
GlaxoSmithKline PLC engages in the creation, discovery, development, manufacture, and marketing of pharmaceutical products, vaccines, over-the-counter medicines, and health-related consumer products in the United Kingdom, the United States, and internationally.

22

Appendix I

Participant Information Sheet



Manchester
Metropolitan
University

Participant Information Sheet

A Smart Data Ecosystem for the Monitoring of Financial Market Irregularities

1. Invitation to research

My name is Lewis Evans, I am completing my PhD at Manchester Metropolitan University. I am conducting this survey to evalauate a financial market monitoring system which has been developed as part of this PhD project.

We would like to invite you to take part in evaluating the effectiveness of a financial market monitoring system (herein named the ‘ecosystem’).

This research project has designed and developed an ecosystem capable of monitoring the financial market – principally discussions surrounding stocks.

2. Why have I been invited?

You have been invited to participate as you have been identified as having sufficient knowledge regarding stock markets and their operation.

3. Do I have to take part?

It is up to you to decide. We will describe the study and go through the information sheet, which we will give to you. We will then ask you to sign a consent form to show you agreed to take part. You are free to withdraw at any time, without giving a reason.

4. What will I be asked to do?

You will be invited to attend a one-to-one meeting, which will take place virtually over Microsoft Teams.

This meeting is expected to last around **forty-five minutes**. The interview will introduce you to a financial market monitoring system, which will be referred to as the ‘ecosystem’. This ecosystem has been developed to monitor financial market irregularities.

During this interview, you will be shown **six** different scenarios that the ecosystem has considered as being potentially irregular. You will be asked questions during the interview relating to these detection scenarios. The investigator will take notes throughout the meeting, which will summarise your thoughts and responses to questions. You will be given the opportunity to review these notes with the interviewer and make amendments should you wish to do so. You will then sign off on these notes to say you are happy with the summary. A summary of these notes will be anonaymised and included in the Evaluation chapter of my PhD thesis and future publications.

5. Are there any risks if I participate?

There are no known risks associated with participating in this study.

6. Are there any advantages if I participate?

There will be no monetary gain from taking part. Instead, you will be contributing to evaluating the effectiveness of a financial market monitoring system, which could assist in the future development of monitoring financial market irregularities. This could assist in making financial markets more transparent.

7. What will happen with the data I provide?

When you agree to participate in this research, we will collect from you personally-identifiable information. Only the PI will have access to personal identificable data (through the consent form). The summary of your thoughts will be anonymised in any research output (e.g. Participant A). All data will be destroyed within 12 months of project completion.

The Manchester Metropolitan University ('the University') is the Data Controller in respect of this research and any personal data that you provide as a research participant.

The University is registered with the Information Commissioner's Office (ICO), and manages personal data in accordance with the General Data Protection Regulation (GDPR) and the University's Data Protection Policy.

We collect personal data as part of this research (such as name, telephone numbers or age). As a public authority acting in the public interest we rely upon the 'public task' lawful basis. When we collect special category data (such as medical information or ethnicity) we rely upon the research and archiving purposes in the public interest lawful basis.

Your rights to access, change or move your information are limited, as we need to manage your information in specific ways in order for the research to be reliable and accurate. If you withdraw from the study, we will keep the information about you that we have already obtained.

We will not share your personal data collected in this form with any third parties.

If your data is shared this will be under the terms of a Research Collaboration Agreement which defines use, and agrees confidentiality and information security provisions. It is the University's policy to only publish anonymised data unless you have given your explicit written consent to be identified in the research. **The University never sells personal data to third parties.**

We will only retain your personal data for as long as is necessary to achieve the research purpose.

For further information about use of your personal data and your data protection rights please see the University's Data Protection Pages.

9. What will happen to the results of the research study?

The research study results will be summarised in the researcher's doctoral thesis to evaluate the ecosystem's effectiveness in flagging potentially irregular behaviour. The results may also be utilised in future journal and/or conference publications.

10. Who has reviewed this research project?

Academic supervisors of the principal investigator have reviewed this research project. In addition, work relating to this ecosystem has previously been reviewed in peer-reviewed journals.

11. Who do I contact if I have concerns about this study or I wish to complain?

Researcher (for general questions):
Lewis Evans
l.evans@mmu.ac.uk
Manchester Metropolitan University
M1 5GD UK
Manchester
England
United Kingdom

Supervisor (for general questions):
Professor Keeley Crockett
k.crockett@mmu.ac.uk
01612471497
Manchester Metropolitan University
M1 5GD UK
Manchester
England
United Kingdom

Faculty Ethics Contact (concerns/complaints):
Science and Engineering Ethics
ethics-scieng@mmu.ac.uk

Manchester Metropolitan University
M1 5GD UK
Manchester
England
United Kingdom

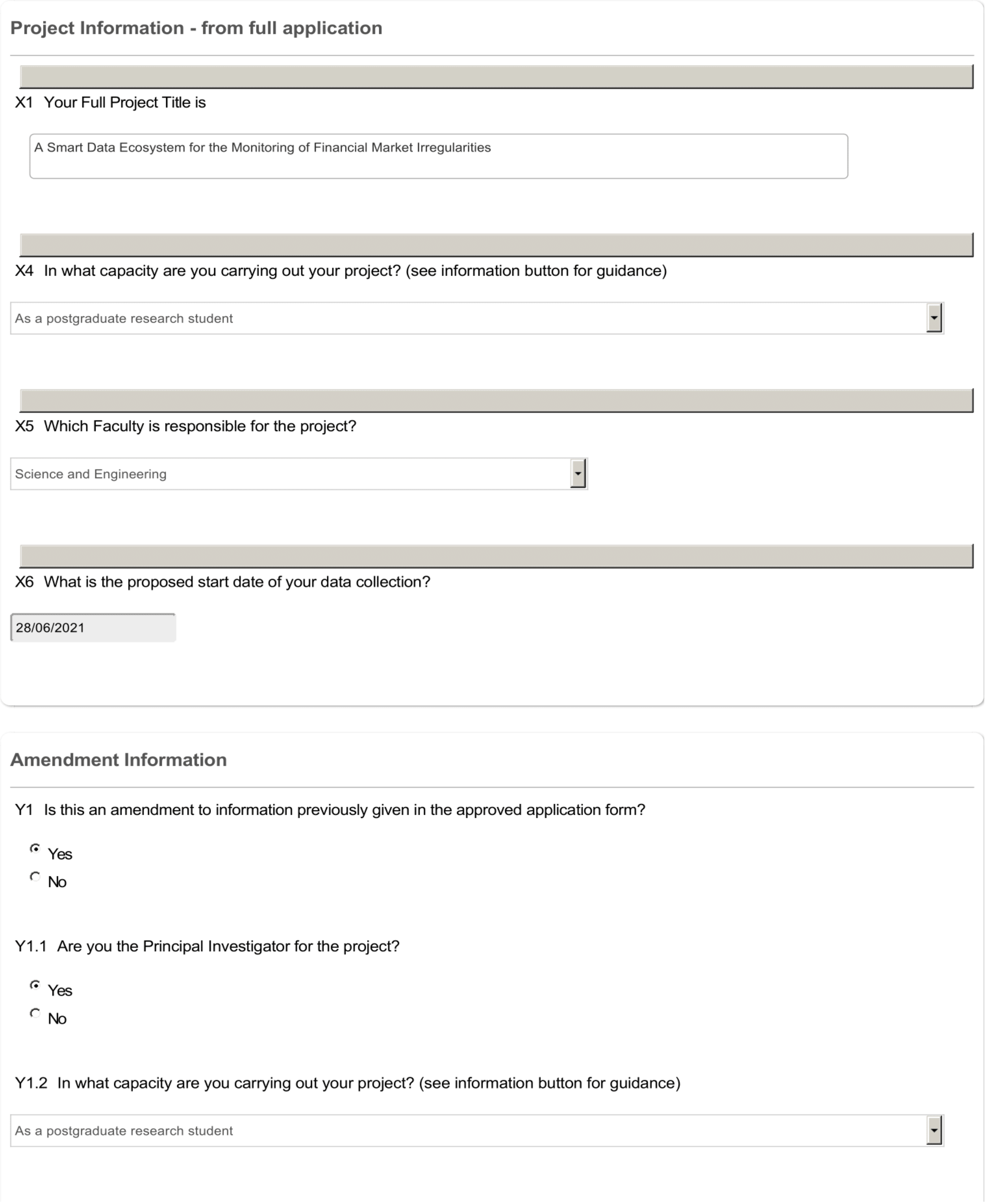
DPO / ICO (concerns relaying to personal data collection):

If you have any concerns regarding the personal data collected from you, our Data Protection Officer can be contacted using the legal@mmu.ac.uk e-mail address, by calling 0161 247 3331 or in writing to: Data Protection Officer, Legal Services, All Saints Building, Manchester Metropolitan University, Manchester, M15 6BH. You also have a right to lodge a complaint in respect of the processing of your personal data with the Information Commissioner's Office as the supervisory authority. Please see: https://ico.org.uk/global/contact-us/

THANK YOU FOR CONSIDERING PARTICIPATING IN THIS PROJECT

Appendix J

System Evaluation Ethical Approval



Project Information - from full application

X1 Your Full Project Title is

A Smart Data Ecosystem for the Monitoring of Financial Market Irregularities

X4 In what capacity are you carrying out your project? (see information button for guidance)

As a postgraduate research student

X5 Which Faculty is responsible for the project?

Science and Engineering

X6 What is the proposed start date of your data collection?

28/06/2021

Amendment Information

Y1 Is this an amendment to information previously given in the approved application form?

● Yes

○ No

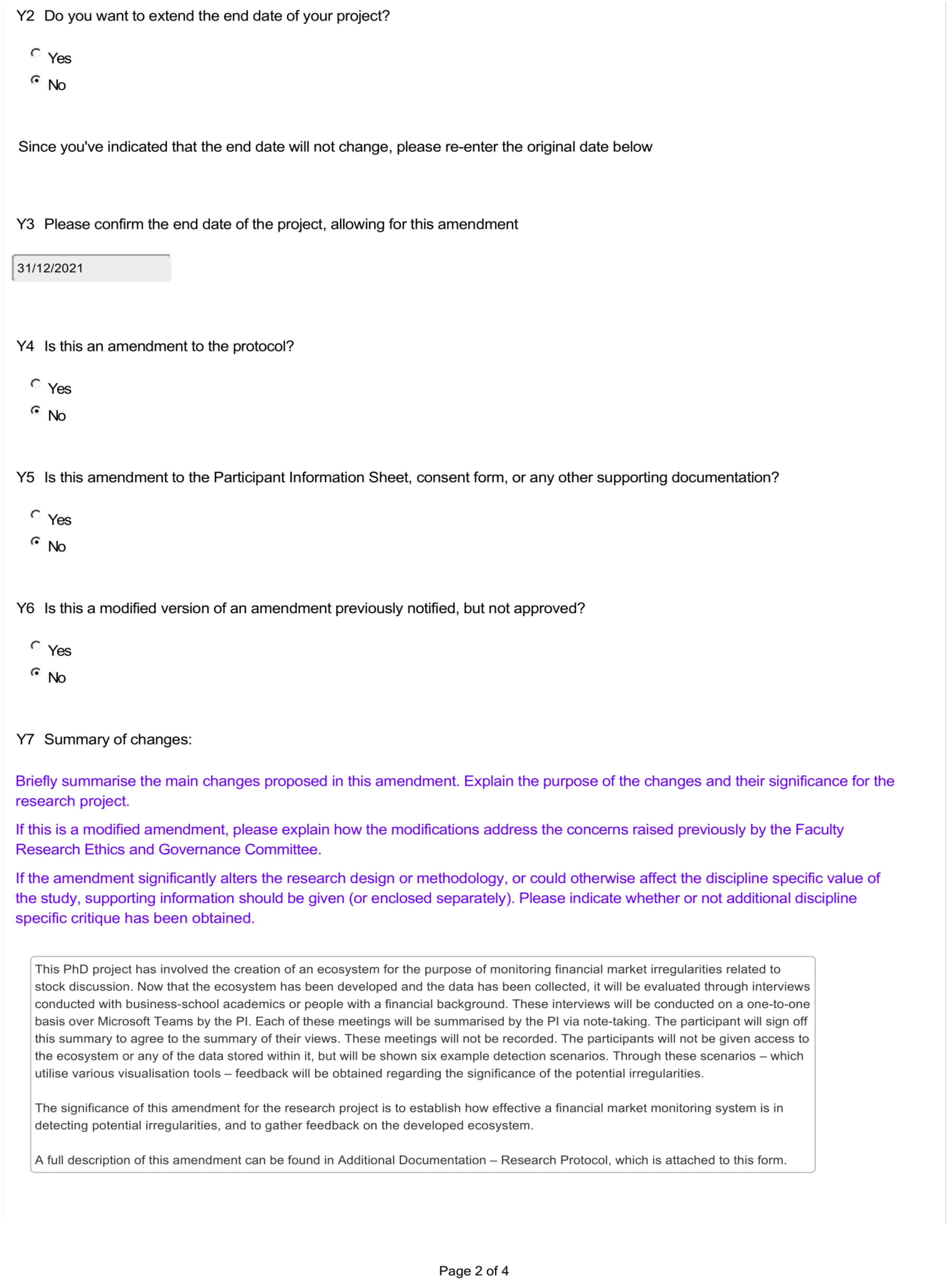
Y1.1 Are you the Principal Investigator for the project?

● Yes

○ No

Y1.2 In what capacity are you carrying out your project? (see information button for guidance)

As a postgraduate research student



Y2 Do you want to extend the end date of your project?

○ Yes
● No

Since you've indicated that the end date will not change, please re-enter the original date below

Y3 Please confirm the end date of the project, allowing for this amendment

31/12/2021

Y4 Is this an amendment to the protocol?

○ Yes
● No

Y5 Is this amendment to the Participant Information Sheet, consent form, or any other supporting documentation?

○ Yes
● No

Y6 Is this a modified version of an amendment previously notified, but not approved?

○ Yes
● No

Y7 Summary of changes:

Briefly summarise the main changes proposed in this amendment. Explain the purpose of the changes and their significance for the research project.

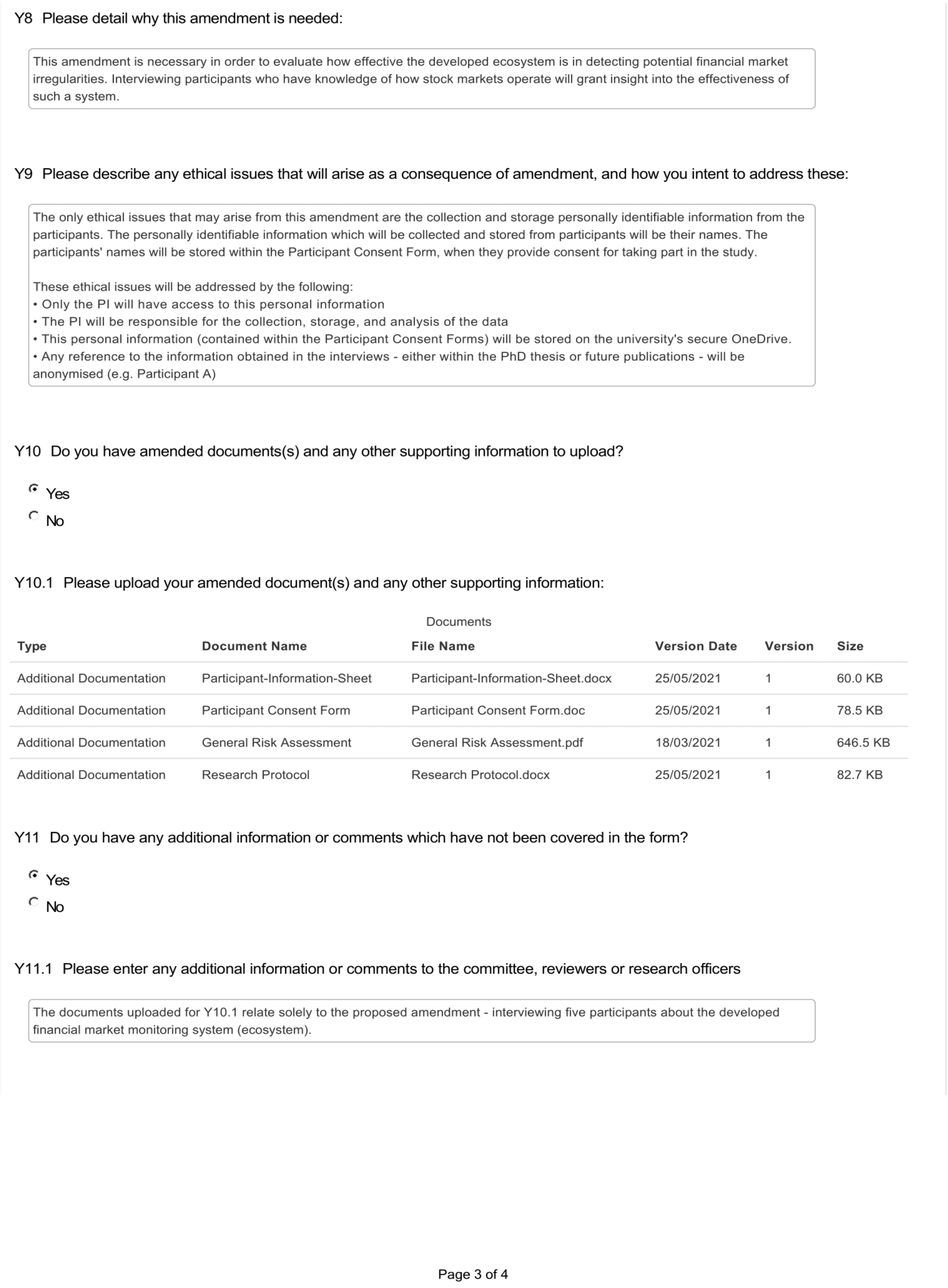
If this is a modified amendment, please explain how the modifications address the concerns raised previously by the Faculty Research Ethics and Governance Committee.

If the amendment significantly alters the research design or methodology, or could otherwise affect the discipline specific value of the study, supporting information should be given (or enclosed separately). Please indicate whether or not additional discipline specific critique has been obtained.

This PhD project has involved the creation of an ecosystem for the purpose of monitoring financial market irregularities related to stock discussion. Now that the ecosystem has been developed and the data has been collected, it will be evaluated through interviews conducted with business-school academics or people with a financial background. These interviews will be conducted on a one-to-one basis over Microsoft Teams by the PI. Each of these meetings will be summarised by the PI via note-taking. The participant will sign off this summary to agree to the summary of their views. These meetings will not be recorded. The participants will not be given access to the ecosystem or any of the data stored within it, but will be shown six example detection scenarios. Through these scenarios – which utilise various visualisation tools – feedback will be obtained regarding the significance of the potential irregularities.

The significance of this amendment for the research project is to establish how effective a financial market monitoring system is in detecting potential irregularities, and to gather feedback on the developed ecosystem.

A full description of this amendment can be found in Additional Documentation – Research Protocol, which is attached to this form.



Y8 Please detail why this amendment is needed:

This amendment is necessary in order to evaluate how effective the developed ecosystem is in detecting potential financial market irregularities. Interviewing participants who have knowledge of how stock markets operate will grant insight into the effectiveness of such a system.

Y9 Please describe any ethical issues that will arise as a consequence of amendment, and how you intent to address these:

The only ethical issues that may arise from this amendment are the collection and storage personally identifiable information from the participants. The personally identifiable information which will be collected and stored from participants will be their names. The participants' names will be stored within the Participant Consent Form, when they provide consent for taking part in the study.

These ethical issues will be addressed by the following:
- Only the PI will have access to this personal information
- The PI will be responsible for the collection, storage, and analysis of the data
- This personal information (contained within the Participant Consent Forms) will be stored on the university's secure OneDrive.
- Any reference to the information obtained in the interviews - either within the PhD thesis or future publications - will be anonymised (e.g. Participant A)

Y10 Do you have amended documents(s) and any other supporting information to upload?

● Yes
○ No

Y10.1 Please upload your amended document(s) and any other supporting information:

Documents

Type	Document Name	File Name	Version Date	Version	Size
Additional Documentation	Participant-Information-Sheet	Participant-Information-Sheet.docx	25/05/2021	1	60.0 KB
Additional Documentation	Participant Consent Form	Participant Consent Form.doc	25/05/2021	1	78.5 KB
Additional Documentation	General Risk Assessment	General Risk Assessment.pdf	18/03/2021	1	646.5 KB
Additional Documentation	Research Protocol	Research Protocol.docx	25/05/2021	1	82.7 KB

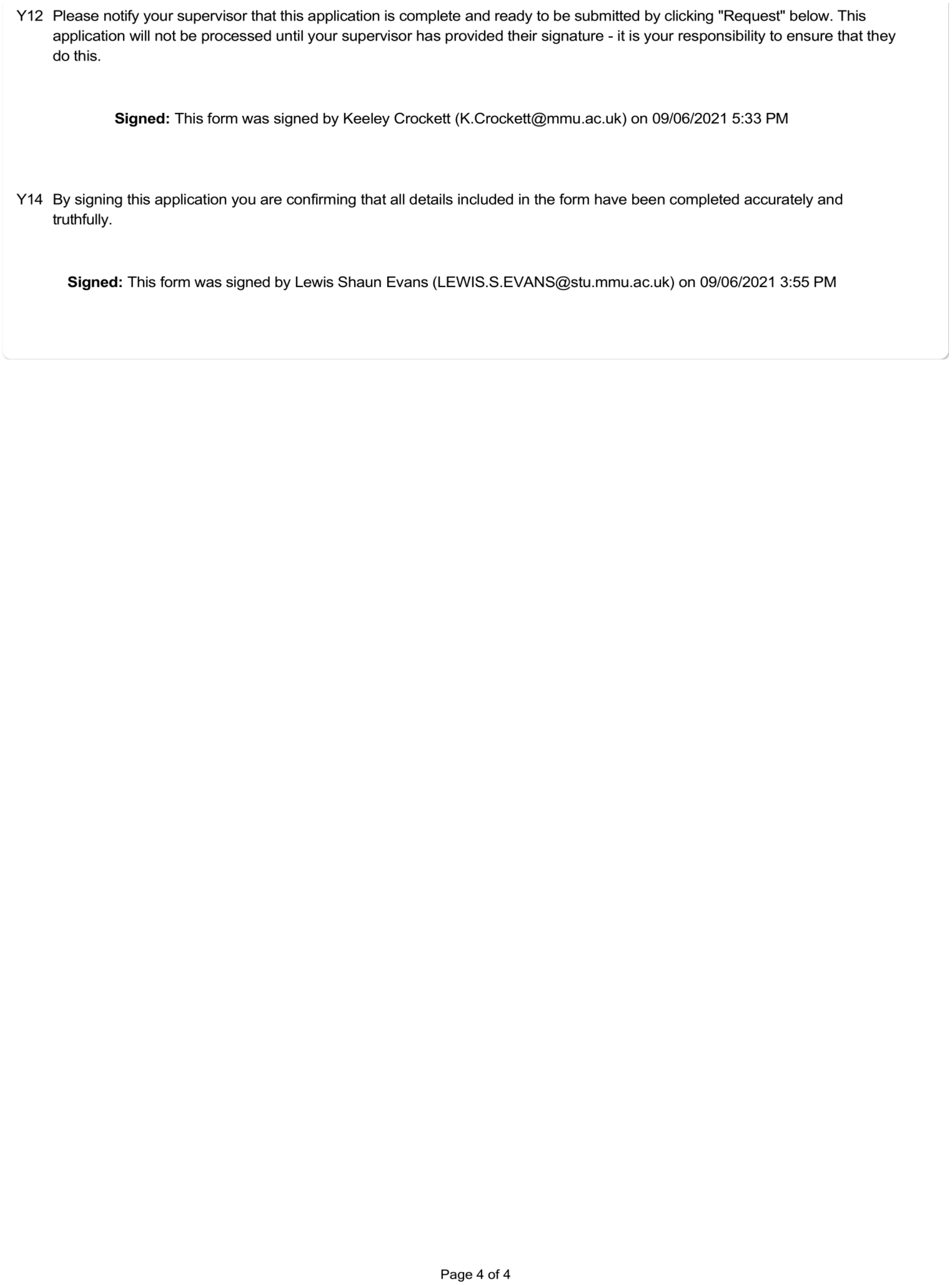
Y11 Do you have any additional information or comments which have not been covered in the form?

● Yes
○ No

Y11.1 Please enter any additional information or comments to the committee, reviewers or research officers

The documents uploaded for Y10.1 relate solely to the proposed amendment - interviewing five participants about the developed financial market monitoring system (ecosystem).

Page 3 of 4



Y12 Please notify your supervisor that this application is complete and ready to be submitted by clicking "Request" below. This application will not be processed until your supervisor has provided their signature - it is your responsibility to ensure that they do this.

Signed: This form was signed by Keeley Crockett (K.Crockett@mmu.ac.uk) on 09/06/2021 5:33 PM

Y14 By signing this application you are confirming that all details included in the form have been completed accurately and truthfully.

Signed: This form was signed by Lewis Shaun Evans (LEWIS.S.EVANS@stu.mmu.ac.uk) on 09/06/2021 3:55 PM

Bibliography

Aghdam, Mehdi Hosseinzadeh, Nasser Ghasem-Aghaee, and Mohammad Ehsan Basiri (2009). "Text feature selection using ant colony optimization". In: *Expert systems with applications* 36.3, pp. 6843–6853.

Ahmed, Mohiuddin, Abdun Naser Mahmood, and Md Rafiqul Islam (2016). "A survey of anomaly detection techniques in financial domain". In: *Future Generation Computer Systems* 55, pp. 278–288. ISSN: 0167739X. DOI: 10.1016/j.future.2015.01.001. URL: http://dx.doi.org/10.1016/j.future.2015.01.001.

Aldeco-Pérez, Rocío and Luc Moreau (2008). "Provenance-based Auditing of Private Data Use". In: DOI: 10.14236/ewic/vocs2008.13.

Alić, Irina (2015). "Supporting Financial Market Surveillance: An IT Artifact Evaluation". In: *28th Bled eConference*, pp. 16–31.

Alnajran, Noufa Abdulaziz and Digital Technology (2019). "An Integrated Semantic-Based Framework for Intelligent Similarity Measurement and Clustering of Microblogging Posts Noufa Abdulaziz Alnajran A thesis submitted in partial fulfilment of the requirements of the Manchester Metropolitan University for the deg". PhD thesis. Manchester Metropolitan University.

Alroobaea, Roobaea and Pam J Mayhew (2014). "How many participants are really enough for usability studies?" In: *Proceedings of 2014 Science and Information Conference, SAI 2014*. Vol. 48, pp. 48–56. ISBN: 9780989319317. DOI: 10.1109/SAI.2014.6918171. URL: www.conference.thesai.org.

Alrubaian, Majed et al. (2018). "Credibility in Online Social Networks: A Survey". In: *IEEE Access* 7, pp. 2828–2855. DOI: 10.1109/ACCESS.2018.2886314. URL: http://www.ieee.org/publications%7B%5C_%7Dstandards/publications/rights/index.html.

Alyannezhadi, Mohammad M., Ali A. Pouyan, and Vahid Abolghasemi (2017). "An efficient algorithm for multisensory data fusion under uncertainty condition". In: *Journal of Electrical Systems and Information Technology* 4.1, pp. 269–278. ISSN: 23147172. DOI: 10.1016/j.jesit.2016.08.002. URL: http://linkinghub.elsevier.com/retrieve/pii/S2314717216300630.

Ambusaidi, Mohammed A et al. (2016). "Building an intrusion detection system using a filter-based feature selection algorithm". In: *IEEE transactions on computers* 65.10, pp. 2986–2998.

Arauzo-Azofra, Antonio, José Luis Aznarte, and José M Benítez (2011). "Empirical study of feature selection methods based on individual feature evaluation for classification problems". In: *Expert systems with applications* 38.7, pp. 8170–8177.

Arnold, Glen (2014). *The Financial Times Guide to Investing*. 3rd. Pearson. ISBN: 9780273723745.

Asghar, Muhammad Zubair et al. (2014). "A review of feature extraction in sentiment analysis". In: *Journal of Basic and Applied Scientific Research* 4.3, pp. 181–186.

Avalon, Grant et al. (2017). "Multi-factor Statistical Arbitrage Model". URL: https://web.stanford.edu/class/msande448/2017/Final/Reports/gr6.pdf.

Bhattacharya, Sanmitra et al. (2012). "Belief surveillance with twitter". In: *Proceedings of the 4th Annual ACM Web Science Conference*, pp. 43–46.

Bholowalia, Purnima and Arvind Kumar (2014). “EBK-means: A clustering technique based on elbow method and k-means in WSN”. In: *International Journal of Computer Applications* 105.9.

Blasch, Erik et al. (2013). “Revisiting the JDL model for information exploitation”. In: *Proceedings of the 16th International Conference on Information Fusion, FUSION 2013*, pp. 129–136. ISBN: 978-605863111-3. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84890807886%7B%5C&%7DpartnerID=40%7B%5C&%7Dmd5=78b887a1e83b6f61bfb80450c5cfa592%7B%5C%%7D5Cnhttp://www.scopus.com/inward/record.url?eid=2-s2.0-84890807886%7B%5C&%7DpartnerID=tZOtx3y1%7B%5C%%7D5Cnhttp://www.scopus.com/inward/record.url?eid=2-s2.0-848.

Bleiholder, Jens and Felix Naumann (2008). “Data fusion”. In: *ACM Computing Surveys* 41.1, pp. 1–41. ISSN: 03600300. DOI: 10.1145/1456650.1456651. URL: http://portal.acm.org/citation.cfm?doid=1456650.1456651.

Bommert, Andrea et al. (2020). “Benchmark for filter methods for feature selection in high-dimensional classification data”. In: *Computational Statistics & Data Analysis* 143, p. 106839.

Bosch, Jan and Petra M. Bosch-Sijtsema (2010). “Softwares product lines, global development and ecosystems: Collaboration in software engineering”. In: *Collaborative Software Engineering*. Springer Berlin Heidelberg, pp. 77–92. ISBN: 9783642102936. DOI: 10.1007/978-3-642-10294-3_4. URL: https://link.springer.com/chapter/10.1007/978-3-642-10294-3%7B%5C_%7D4.

Bountouridis, Dimitrios et al. (2019). “Annotating credibility: Identifying and mitigating bias in credibility datasets”. In: *ROME 2019-Workshop on Reducing Online Misinformation Exposure*.

Bozdogan, Hamparsum (1987). “Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions”. In: *Psychometrika* 52.3, pp. 345–370.

Byrka-Kita, Katarzyna, Mateusz Czerwiński, and Agnieszka Preś-Perepeczo (2017). “Stock market reaction to CEO appointment–preliminary results”. In: *Central European Management Journal* 25.2, pp. 23–42.

Campbell, John and Byron Keating (2013). “Development of a decision support system for collective stock market intelligence”. In: *Proceedings of the 24th Australasian Conference on Information Systems*. ISBN: 9780992449506. URL: https://aisel.aisnet.org/acis2013/93.

Castanedo, F (2013). “A review of data fusion techniques”. In: *ScientificWorldJournal* 2013, p. 704504. ISSN: 1537-744X. DOI: 10.1155/2013/704504. URL: http://www.ncbi.nlm.nih.gov/pubmed/24288502.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete (2011). “Information Credibility on Twitter”. In: *Proceedings of the 20th international conference on World Wide Web*, pp. 675–684. ISBN: 9781450306324. URL: http://blog.twitter.com/2010/09/evolving-ecosystem.html.

Castillo, Samara et al. (2019). “Detection of Bots and Cyborgs in Twitter: A study on the Chilean Presidential Election in 2017”. In: *International Conference on Human-Computer Interaction*. Springer, pp. 311–323.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar (2009). *Anomaly detection: A survey*. DOI: 10.1145/1541880.1541882.

Chaturvedi, Anil et al. (July 2001). “K-modes clustering”. In: *Journal of Classification* 18.1, pp. 35–55. ISSN: 01764268. DOI: 10.1007/s00357-001-0004-3. URL: https://link.springer.com/article/10.1007/s00357-001-0004-3.

Chawla, Nitesh V et al. (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

Chen, Kaiping, Zening Duan, and Sijia Yang (2021). "Twitter as research data: Tools, costs, skill sets, and lessons learned". In: *Politics and the Life Sciences*, pp. 1–17. ISSN: 0730-9384. DOI: 10.1017/PLS.2021.19. URL: https://www.cambridge.org/core/journals/politics-and-the-life-sciences/article/twitter-as-research-data/6B31D18C5E2F9B8F9C0301BFB05F1C27.

Chicco, Davide and Giuseppe Jurman (2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC genomics* 21.1, pp. 1–13.

Clemons, Eric K and Jennifer T Adams (1988). "International opportunities for the use of information systems in securities trading created by deregulation of the London Stock Exchange". In: *Office Technology and People* 4.4, pp. 271–284. ISSN: 0167-5710. DOI: 10.1108/eb022666.

Close, Liam and Rasha Kashef (2020). "Combining Artificial Immune System and Clustering Analysis: A Stock Market Anomaly Detection Model". In: *Journal of Intelligent Learning Systems and Applications* 12.04, pp. 83–108. ISSN: 2150-8402. DOI: 10.4236/jilsa.2020.124005. URL: https://doi.org/10.4236/jilsa.2020.124005.

Cresci, Stefano, Roberto Di Pietro, et al. (2015). "Fame for sale: Efficient detection of fake Twitter followers". In: *Decision Support Systems* 80, pp. 56–71.

Cresci, Stefano, Fabrizio Lillo, et al. (2018). "Cashtag piggybacking: uncovering spam and bot activity in stock microblogs on Twitter". In: *ACM Transactions on the Web* 13.2, pp. 1–18. arXiv: 1804.04406. URL: http://arxiv.org/abs/1804.04406.

Cuesta-Albertos, J. A., A. Gordaliza, and C. Matrán (Apr. 1997). "Trimmed k-means: An attempt to robustify quantizers". In: *Annals of Statistics* 25.2, pp. 553–576. ISSN: 00905364. DOI: 10.1214/aos/1031833664. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-25/issue-2/Trimmed-k-means-an-attempt-to-robustify-quantizers/10.1214/aos/1031833664.full%20https://projecteuclid.org/journals/annals-of-statistics/volume-25/issue-2/Trimmed-k-means-an-attempt-t.

Dasarathy, Belur V (1997). "Sensor fusion potential exploitation-innovative architectures and illustrative applications". In: *Proceedings of the IEEE*. Vol. 85. 1, pp. 24–38. DOI: 10.1109/5.554206.

De Micheli, Carlo and Andrea Stroppa (2013). "Twitter and the underground market". In: *11th Nexa Lunch Seminar*. Vol. 22.

Diaz, David et al. (2011). "A Systematic framework for the analysis and development of financial market monitoring systems". In: *Proceedings - 2011 Annual SRII Global Conference, SRII 2011*. March 2014, pp. 145–153. ISBN: 9780769543710. DOI: 10.1109/SRII.2011.27.

Dorado, Hugo et al. (2019). "Wrapper for building classification models using Covering Arrays". In: *IEEE Access* 7, pp. 148297–148312.

Doukas, John A and Hafiz Hoque (2016). "Why firms favour the AIM when they can list on main market?" In: *Journal of International Money and Finance* 60, pp. 378–404. ISSN: 0261-5606. DOI: 10.1016/j.jimonfin.2015.10.001. URL: http://dx.doi.org/10.1016/j.jimonfin.2015.10.001.

Duong, Trong H, Hong Q Nguyen, and Geun S Jo (2017). "Smart Data : Where the Big Data Meets the Semantics". In: *Computational Intelligence and Neuroscience* 2, pp. 1–2. ISSN: 16875273. DOI: 10.1155/2017/6925138.

Eisler, Zoltán, János Kertész, and Fabrizio Lillo (2007). "The limit order book on different time scales". In: *Noise and Stochastics in Complex Systems and Finance*.

Vol. 6601, pp. 1–11. ISBN: 0819467383. DOI: 10.1117/12.724817. arXiv: 0705.4023.

El Ballouli, Rim et al. (2017). "Cat: Credibility analysis of arabic content on twitter". In: *Proceedings of the Third Arabic Natural Language Processing Workshop*, pp. 62–71.

Evans, Lewis, Majdi Owda, Keeley Crockett, and Ana Fernández Vilas (2018). "Big Data Fusion Model for Heterogeneous Financial Market Data (FinDF)". In: *Proceedings of the 2018 Intelligent Systems Conference (IntelliSys)*. September, pp. 1085–1101.

Evans, Lewis, Majdi Owda, Keeley Crockett, and Ana Fernandez Vilas (Aug. 2019). "A methodology for the resolution of cashtag collisions on Twitter – A natural language processing & data fusion approach". In: *Expert Systems with Applications* 127, pp. 353–369. ISSN: 09574174. DOI: 10.1016/j.eswa.2019.03.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417419301812.

— (Apr. 2021). "Credibility assessment of financial stock tweets". In: *Expert Systems with Applications* 168, p. 114351. ISSN: 09574174. DOI: 10.1016/j.eswa.2020.114351.

Faber, Vance (1994). "Clustering and the Continuous k-Means Algorithm". In: *Los Alamos Science* 22.22, pp. 138–144. URL: https://staff.fmi.uvt.ro/%7B~%7Ddaniela.zaharie/dm2019/RO/proiecte/biblio/kMeans/ContinuousKMeans.pdf.

Fama, Eugene F. (May 1970). "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *The Journal of Finance* 25.2, p. 383. ISSN: 00221082. DOI: 10.2307/2325486.

Fernández Vilas, Ana et al. (2017). "Experiment for analysing the impact of financial events on twitter". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10393 LNCS. Springer Verlag, pp. 407–419. ISBN: 9783319654812. DOI: 10.1007/978-3-319-65482-9_28. URL: https://link.springer.com/chapter/10.1007/978-3-319-65482-9%7B%5C_%7D28.

Financial Conduct Authority (Apr. 2019). *Enforcement — FCA*. URL: https://www.fca.org.uk/about/enforcement (visited on 04/21/2021).

— (2021). *Financial Conduct Authority Fines*. URL: https://www.fca.org.uk/news/news-stories/2021-fines (visited on).

Financial Services Act 2012 (Dec. 2012). *Financial Services Act 2012*. URL: https://www.legislation.gov.uk/ukpga/2012/21/2013-04-12%7B%5C#%7DenTabHelp.

Flood, Mark, H Jagadish, and Louiqa Raschid D (2016). "Big data challenges and opportunities in financial stability monitoring". In: *Financial Stability Review* 20, pp. 1–20.

Fowler, Martin (2004). *UML distilled: a brief guide to the standard object modeling language*. Addison-Wesley Professional.

Freund, Yoav and Llew Mason (1999). "The alternating decision tree learning algorithm". In: *icml*. Vol. 99. Citeseer, pp. 124–133.

FTSE Russell (2021). *Industry Classification Benchmark (ICB) — FTSE Russell*. URL: https://www.ftserussell.com/data/industry-classification-benchmark-icb (visited on 05/03/2021).

Géron, Aurélien (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. Ed. by Rachel Roumeliotis and Nicole Tache. 2nd. O'Reilly.

Goldberg, Henry G et al. (2003). "The NASD Securities Observation, New Analysis and Regulation System (SONAR)". In: *Iaai*, pp. 11–18. URL: https://pdfs.semanticscholar.org/1a20/2ce3b60fc7d9a758c15bc6a5d66eb394b390.pdf.

Golmohammadi, Koosha et al. (2015). “Time Series Contextual Anomaly Detection for Detecting Market Manipulation in Stock Market”. In: *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. ISBN: 9781467382731. DOI: `10.1109/DSAA.2015.7344856`.

Gondhalekar, Vijay and Sonia Dalmia (2007). “Examining the stock market response: a comparison of male and female CEOs”. In: *International Advances in Economic Research* 13.3, pp. 395–397.

Gorrell, Genevieve, Johann Petrak, and Kalina Bontcheva (2015). “Using@ Twitter conventions to improve# lod-based named entity disambiguation”. In: *European semantic web conference*. Springer, pp. 171–186.

Gupta, Aditi and Ponnurangam Kumaraguru (2012). “Credibility ranking of tweets during high impact events”. In: *Proceedings of the 1st workshop on privacy and security in online social media*, pp. 2–8.

Gupta, Aditi, Ponnurangam Kumaraguru, et al. (2014). “Tweetcred: Real-time credibility assessment of content on twitter”. In: *International conference on social informatics*. Springer, pp. 228–243.

Hall, David L. and James Llinas (1998). “An introduction to multi-sensor data fusion”. In: *ISCAS’98. Proceedings of the 1998 IEEE International Symposium on Circuits and Systems (Cat. No. 98CH36187)*. Vol. 6. 1, pp. 537–540. ISBN: 9781420003161.

Hassan, Noha Y et al. (2018). “Supervised learning approach for twitter credibility detection”. In: *2018 13th International conference on computer engineering and systems (ICCES)*. IEEE, pp. 196–201.

Hayward, C and A. Madill (2004). “A Survey of Outlier Detection Methodologies”. In: *Artificial Intelligence Review* 22.2, pp. 85–126.

Hsu, Hui-Huang, Cheng-Wei Hsieh, and Ming-Da Lu (2011). “Hybrid feature selection by combining filters and wrappers”. In: *Expert Systems with Applications* 38.7, pp. 8144–8150.

Hsueh, Pei-Yun, Prem Melville, and Vikas Sindhwani (2009). “Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria”. In: *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pp. 27–35.

Huang, Zhexue (1998). “Extensions to the k-means algorithm for clustering large data sets with categorical values”. In: *Data Mining and Knowledge Discovery* 2.3, pp. 283–304. ISSN: 13845810. DOI: `10.1023/A:1009769707641`. URL: `https://link.springer.com/article/10.1023/A:1009769707641`.

Inkpen, Diana et al. (Oct. 2017). “Location detection and disambiguation from twitter messages”. In: *Journal of Intelligent Information Systems* 49.2, pp. 237–253. ISSN: 15737675. DOI: `10.1007/s10844-017-0458-3`.

Istiake Sunny, Md Arif, Mirza Mohd Shahriar Maswood, and Abdullah G Alharbi (2020). “Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model”. In: *2nd Novel Intelligent and Leading Emerging Sciences Conference, NILES 2020*, pp. 87–92. ISBN: 9781728182261. DOI: `10.1109/NILES50944.2020.9257950`. URL: `https://www.researchgate.net/publication/347044815`.

Joshi, Ankur et al. (2015). “Likert scale: Explored and explained”. In: *British Journal of Applied Science & Technology* 7.4, p. 396.

Jureviciene, Daiva and Kristina Jermakova (2012). “The Impact of Individuals’ Financial Behaviour on Investment Decisions”. In: *Electronic International Interdisciplinary Conference*, pp. 242–250.

Kaloyanova, Elitsa (2020). *How to Combine PCA and K-Means Clustering in Python?* URL: `https://365datascience.com/tutorials/python-tutorials/pca-k-means/` (visited on 08/16/2022).

Khaleghi, Bahador et al. (2013). "Multisensor data fusion : A review of the state-of-the-art". In: *Information Fusion* 14.1, pp. 28–44. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2011.08.001. URL: http://dx.doi.org/10.1016/j.inffus.2011.08.001.

Khan, Nawsher et al. (2019). "The 51 V's of big data: Survey, technologies, characteristics, opportunities, issues and challenges". In: *ACM International Conference Proceeding Series*. Vol. Part F1481, pp. 19–24. DOI: 10.1145/3312614.3312623.

Kim, Sang-Bum et al. (2006). "Some effective techniques for naive bayes text classification". In: *IEEE transactions on knowledge and data engineering* 18.11, pp. 1457–1466.

Kim, Yoonseong and So Young Sohn (2012). "Stock fraud detection using Peer Group Analysis". In: *Expert Systems With Applications* 39.10, pp. 8986–8992. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2012.02.025. URL: http://dx.doi.org/10.1016/j.eswa.2012.02.025.

Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data". In: *biometrics*, pp. 159–174.

Laney, Doug (2001). "3D data management: Controlling data volume, velocity and variety". In: *Application Delivery Strategies* 949.February 2001, p. 4. URL: https://scholar.google.com/scholar%7B%5C_%7Dlookup?title=3d%20Data%20managment%7B%5C%%7D3A%20controlling%20data%20volume%7B%5C%%7D2C%20velocity%20and%20variety%7B%5C&%7Dauthor=Doug%20Laney%7B%5C&%7Dpublication%7B%5C_%7Dyear=2001%20http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Lee, Pei Shyuan, Majdi Owda, and Keeley Crockett (2018). "The detection of fraud activities on the stock market through forward analysis methodology of financial discussion boards". In: *Future of Information and Communications Conference (FICC) 2018 5-6 April 2018 — Singapore* April.

Leković, Miljan (2018). "Evidence for and against the validity of efficient market hypothesis". In: *Economic themes* 56.3, pp. 369–387.

Li, Guoquan et al. (2018). "Data Fusion for Network Intrusion Detection: A Review". In: *Security and Communication Networks* 2018. ISSN: 19390122. DOI: 10.1155/2018/8210614.

Li, Qing et al. (2014). "The effect of news and public mood on stock movements". In: *Information Sciences* 278, pp. 826–840. ISSN: 0020-0255. DOI: 10.1016/j.ins.2014.03.096. URL: http://dx.doi.org/10.1016/j.ins.2014.03.096.

Li, Xiaodong, Xiaodi Huang, et al. (2014). "Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information". In: *Neurocomputing* 142, pp. 228–238. ISSN: 18728286. DOI: 10.1016/j.neucom.2014.04.043. URL: http://dx.doi.org/10.1016/j.neucom.2014.04.043.

Li, Xiaodong, Haoran Xie, et al. (Oct. 2014). "News impact on stock price return via sentiment analysis". In: *Knowledge-Based Systems* 69, pp. 14–23. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2014.04.022. URL: http://dx.doi.org/10.1016/j.knosys.2014.04.022%20https://linkinghub.elsevier.com/retrieve/pii/S0950705114001440.

Liu, Yingbo et al. (2015). "Research on the Matthews correlation coefficients metrics of personalized recommendation algorithm evaluation". In: *International Journal of Hybrid Information Technology* 8.1, pp. 163–172. ISSN: 17389968. DOI: 10.14257/ijhit.2015.8.1.14.

London Stock Exchange (2021). *Broker Directory*. URL: https://www.londonstockexchange.com/personal-investing/member-firm-broker-directory (visited on).

Lorek, Krzysztof et al. (2015). "Automated credibility assessment on twitter". In: *Computer Science* 16.2), pp. 157–168.

Loughran, Tim and Bill McDonald (2011). "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks". In: *The Journal of finance* 66.1, pp. 35–65.

— (2016). "Textual analysis in accounting and finance: A survey". In: *Journal of Accounting Research* 54.4, pp. 1187–1230.

Majumdar, Adrija and Indranil Bose (2018). "Detection of financial rumors using big data analytics: the case of the Bombay Stock Exchange". In: *Journal of Organizational Computing and Electronic Commerce* 28.2, pp. 79–97. ISSN: 10919392. DOI: `10.1080/10919392.2018.1444337`. URL: `https://doi.org/10.1080/10919392.2018.1444337`.

Manikas, Konstantinos and Klaus Marius Hansen (2013). "Software ecosystems - A systematic literature review". In: *Journal of Systems and Software* 86.5, pp. 1294–1306. ISSN: 01641212. DOI: `10.1016/j.jss.2012.12.026`. URL: `http://dx.doi.org/10.1016/j.jss.2012.12.026`.

McDaniel, D (2001). "An Information Fusion Framework for Data Integration". In: *Silver Bullet Solutions* 858.

Meng, Tong et al. (2020). "A survey on machine learning for data fusion". In: *Information Fusion* 57.2, pp. 115–129. ISSN: 15662535. DOI: `10.1016/j.inffus.2019.12.001`. URL: `https://doi.org/10.1016/j.inffus.2019.12.001`.

Messerschmitt, David G. and Clemens Szyperski (2005). "Software Ecosystem: Understanding an Indispensable Technology and Industry". In: *MIT Press Books* 1. URL: `https://ideas.repec.org/b/mtp/titles/0262633310.html`.

Mishkin, Frederic (2016). *The Economics of Money, Banking, and Financial Markets*. 11th. Pearson. ISBN: 9788578110796. arXiv: `arXiv:1011.1669v3`.

Mood, Carina (2010). "Logistic regression: Why we cannot do what we think we can do, and what we can do about it". In: *European sociological review* 26.1, pp. 67–82.

Morris, Meredith Ringel et al. (2012). "Tweeting is believing? Understanding microblog credibility perceptions". In: *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pp. 441–450.

Nadeau, David and Satoshi Sekine (2007). "A survey of named entity recognition and classification". In: *Lingvisticae Investigationes* 30.1, pp. 3–26.

Ngai, Eric W.T. et al. (2017). "Big data analytics in electronic markets". In: *Electronic Markets* 27.3, pp. 243–245. ISSN: 14228890. DOI: `10.1007/s12525-017-0261-6`.

Nofer, Michael and Oliver Hinz (2015). "Using twitter to predict the stock market". In: *Business & Information Systems Engineering* 57.4, pp. 229–242.

ODonovan, John et al. (2012). "Credibility in context: An analysis of feature distributions in twitter". In: *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, pp. 293–301.

Oliveira, Nuno, Paulo Cortez, and Nelson Areal (2016). "Stock market sentiment lexicon acquisition using microblogging data and statistical measures". In: *Decision Support Systems* 85, pp. 62–73.

Omar, Norshafarina et al. (2013). "Review of feature selection for solving classification problems". In: *Journal of Information System Research and Innovation* 3, pp. 64–70.

Owda, Majdi, Keeley Crockett, and Pei Shyuan Lee (2017). "Financial Discussion Boards Irregularities Detection System (FDBs-IDS) using Information Extraction". In: *Intelligent Systems*. September, pp. 8–12.

Page, Janis Teruggi and Margaret E Duffy (2018). "What does credibility look like? Tweets and walls in US presidential candidates' visual storytelling". In: *Journal of political Marketing* 17.1, pp. 3–31.

Parmezan, Antonio Rafael Sabino, Huei Diana Lee, and Feng Chung Wu (June 2017). "Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework". In: *Expert Systems with Applications* 75, pp. 1–24. ISSN: 09574174. DOI: 10.1016/j.eswa.2017.01.013.

Pelleg, Dan and Andrew Moore (2015). "X-means: Extending K-means with Efficient Estimation of the Number of Clusters". In: *CEUR Workshop Proceedings* 1542, pp. 33–36. ISSN: 16130073. arXiv: arXiv:1011.1669v3. URL: http://web.cs.dal.ca/%7B~%7Dshepherd/courses/csci6403/clustering/xmeans.pdf%20https://www.cs.cmu.edu/%7B~%7Ddpelleg/download/xmeans.pdf.

Pham, Rebecca and Marcel Ausloos (2020). "Insider trading in the run-up to merger announcements. Before and after the UK's Financial Services Act 2012". In: *International Journal of Finance and Economics* October, pp. 1–13. ISSN: 10991158. DOI: 10.1002/ijfe.2325.

Phillips, Steven J. (2002). "Acceleration of k-means and related clustering algorithms". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 2409. Springer Verlag, pp. 166–177. ISBN: 3540439773. DOI: 10.1007/3-540-45643-0_13.

Polansky, S, M Kulczak, and L Fitzpatrick (2004). "NASD Market surveillance assessment and recommendations. Final report". In: *Achievement of Market Friendly Initiatives and Results Program (AMIR 2.0 Program)*.

Premti, Arjan, Luis Garcia-Feijoo, and Jeff Madura (2017). "Information content of analyst recommendations in the banking industry". In: *International Review of Financial Analysis* 49, pp. 35–47. ISSN: 10575219. DOI: 10.1016/j.irfa.2016.11.005. URL: https://www.sciencedirect.com/science/article/pii/S1057521916301806.

Qi, Jun et al. (Mar. 2020). "An overview of data fusion techniques for Internet of Things enabled physical activity recognition and measure". In: *Information Fusion* 55, pp. 269–280. ISSN: 15662535. DOI: 10.1016/j.inffus.2019.09.002.

Rajesh, Neeraj and Lisa Gandy (2016). "CashTagNN: Using sentiment of tweets with CashTags to predict stock market prices". In: *SITA 2016 - 11th International Conference on Intelligent Systems: Theories and Applications*. ISBN: 9781509057818. DOI: 10.1109/SITA.2016.7772262.

Ranco, Gabriele et al. (2015). "The effects of Twitter sentiment on stock price returns". In: *PloS one* 10.9, e0138441.

Reidsma, Dennis and Rieks op den Akker (2008). "Exploiting 'subjective'annotations". In: *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pp. 8–16.

Richert, Willi (2013). *Building machine learning systems with Python*. Packt Publishing Ltd.

Ronaghan, Stacey (2018). *The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-Learn and Spark. 2018.*

Rong, Miao, Dunwei Gong, and Xiaozhi Gao (2019). "Feature selection and its use in big data: challenges, methods, and trends". In: *IEEE Access* 7, pp. 19709–19725.

Rousseeuw, Peter J (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20, pp. 53–65.

Sabherwal, Sanjiv, Salil K Sarkar, and Ying Zhang (2011). "Do internet stock message boards influence trading? Evidence from heavily discussed stocks with no fundamental news". In: *Journal of Business Finance & Accounting* 38.9-10, pp. 1209–1237.

Schulmerich, Marcus, Yves-Michel Leporcher, and Ching-Hwa Eu (2015). *Applied Asset and Risk Management*. Springer, p. 491. ISBN: 9783642554438.

Schumaker, Robert P. and Hsinchun Chen (2006). "Textual analysis of stock market prediction using financial news articles". In: *Association for Information Systems - 12th Americas Conference On Information Systems, AMCIS 2006* 3, pp. 1422–1430.

Shen, Heng Tao et al. (Feb. 2021). "Heterogeneous data fusion for predicting mild cognitive impairment conversion". In: *Information Fusion* 66, pp. 54–63. ISSN: 15662535. DOI: 10.1016/j.inffus.2020.08.023.

Siering, Michael et al. (2017). "A taxonomy of financial market manipulations: establishing trust and market integrity in the financialized economy through automated fraud detection". In: *Journal of Information Technology*. ISSN: 1466-4437. DOI: 10.1057/s41265-016-0029-z.

Sikdar, Sujoy et al. (2013). "Understanding information credibility on twitter". In: *2013 International Conference on Social Computing*. IEEE, pp. 19–24.

Spina, Damiano, Julio Gonzalo, and Enrique Amigó (2013). "Discovering filter keywords for company name disambiguation in twitter". In: *Expert Systems With Applications* 40.12, pp. 4986–5003. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2013.03.001. URL: http://dx.doi.org/10.1016/j.eswa.2013.03.001.

Suresh, M R (2013). "A Study on Fundamental and Technical Analysis". In: *International Journal of Marketing, Financial Services & Management Research* 2.5, pp. 44–59.

Thakkar, Ankit and Kinjal Chaudhari (2021). "Fusion in stock market prediction: A decade survey on the necessity, recent developments, and potential future directions". In: *Information Fusion* 65.June 2020, pp. 95–107. ISSN: 15662535. DOI: 10.1016/j.inffus.2020.08.019. URL: https://doi.org/10.1016/j.inffus.2020.08.019.

Tsai, Chih-Fong and Yu-Chi Chen (2019). "The optimal combination of feature selection and data discretization: An empirical study". In: *Information Sciences* 505, pp. 282–293. ISSN: 00200255. DOI: 10.1016/j.ins.2019.07.091. URL: https://doi.org/10.1016/j.ins.2019.07.091.

Ullah, Ihsan and Hee Yong Youn (Apr. 2020). "Intelligent Data Fusion for Smart IoT Environment: A Survey". In: *Wireless Personal Communications* 114.1, pp. 409–430. ISSN: 1572834X. DOI: 10.1007/s11277-020-07369-0. URL: https://link.springer.com/article/10.1007/s11277-020-07369-0.

Vadera, Sunil (2010). "CSNL: A cost-sensitive non-linear decision tree algorithm". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 4.2, pp. 1–25.

Valverde-albacete, Francisco J and Carmen Pela (2014). "100% Classification Accuracy Considered Harmful : The Normalized Information Transfer Factor Explains the Accuracy Paradox". In: *PLoS ONE* 9.1, pp. 1–10. DOI: 10.1371/journal.pone.0084217.

Verma, Monika and Sanjeev Sofat (2014). "Techniques to detect spammers in twitter- a survey". In: *International Journal of Computer Applications* 85.10.

Wah, Yap Bee et al. (2018). "Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy". In: *Pertanika Journal of Science and Technology* 26.1, pp. 329–340. ISSN: 22318526. URL: http://www.pertanika.upm.edu.my/Pertanika%20PAPERS/JST%20Vol.%2026%20(1)%20Jan.%202018/21%20JST(S)-0296-2017-3rdProof.pdf.

Weng, Bin, Mohamed A. Ahmed, and Fadel M. Megahed (Aug. 2017). "Stock market one-day ahead movement prediction using disparate data sources". In: *Expert Systems with Applications* 79, pp. 153–163. ISSN: 09574174. DOI: 10.1016/j.eswa.2017.02.041.

Wishart, D. (2003). "k-Means Clustering with Outlier Detection, Mixed Variables and Missing Values". In: *Exploratory Data Analysis in Empirical Research*. Springer, Berlin, Heidelberg, pp. 216–226. DOI: 10.1007/978-3-642-55721-7_23. URL: https://link.springer.com/chapter/10.1007/978-3-642-55721-7%7B%5C_%7D23.

Wold, Svante, Kim Esbensen, and Paul Geladi (1987). "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2.1-3, pp. 37–52.

Xu, Dongkuan and Yingjie Tian (2015). "A Comprehensive Survey of Clustering Algorithms". In: *Annals of Data Science* 2.2, pp. 165–193. ISSN: 2198-5804. DOI: 10.1007/s40745-015-0040-1.

Yang, Fan et al. (2012). "Automatic detection of rumor on sina weibo". In: *Proceedings of the ACM SIGKDD workshop on mining data semantics*, pp. 1–7.

Yang, Jingchao et al. (2019). "A twitter data credibility framework—Hurricane Harvey as a use case". In: *ISPRS International Journal of Geo-Information* 8.3, p. 111.

Yang, Min-Chul and Hae-Chang Rim (2014). "Identifying interesting Twitter contents using topical analysis". In: *Expert Systems with Applications* 41.9, pp. 4330–4336.

Yang, Zhao, Xuezheng Sun, and James W Hardin (2011). "Testing marginal homogeneity in clustered matched-pair data". In: *Journal of Statistical Planning and Inference* 141.3, pp. 1313–1318.

Yu, Hualong, Jun Ni, and Jing Zhao (2013). "ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data". In: *Neurocomputing* 101, pp. 309–318.

Zaki, Mohamed, Babis Theodoulidis, and David Diaz (2019). "Ontology-Driven Framework for Stock Market Monitoring and Surveillance". In: *Handbook of Global Financial Markets*. Ed. by Sabri Boubaker and Duc Khuong Nguyen. World Scientific, pp. 75–103.

Zhang, Li Dong et al. (2013). "A K-harmonic means clustering algorithm based on enhanced differential evolution". In: *Proceedings - 2013 5th Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2013*, pp. 13–16. ISBN: 9780769549323. DOI: 10.1109/ICMTMA.2013.1. URL: https://ieeexplore.ieee.org/abstract/document/6493658/.

Zhang, Liang et al. (Jan. 2013). "Based on information fusion technique with data mining in the application of finance early-warning". In: *Procedia Computer Science*. Vol. 17. Elsevier, pp. 695–703. DOI: 10.1016/j.procs.2013.05.090.

Zhang, Xi et al. (2018). "Improving stock market prediction via heterogeneous information fusion". In: *Knowledge-Based Systems* 143, pp. 236–247. ISSN: 09507051. DOI: 10.1016/j.knosys.2017.12.025. arXiv: 1801.00588. URL: https://doi.org/10.1016/j.knosys.2017.12.025.

Zhang, Yu Dong et al. (Dec. 2020). "Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation". In: *Information Fusion* 64, pp. 149–187. ISSN: 15662535. DOI: 10.1016/j.inffus.2020.07.006. URL: /pmc/articles/PMC7366126/%20/pmc/articles/PMC7366126/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7366126/.

Zhou, Hong Bo and Jun Tao Gao (2014). "Automatic method for determining cluster number based on silhouette coefficient". In: *Advanced Materials Research*. Vol. 951. Trans Tech Publ, pp. 227–230.