Novel methods for resolving false positives during the detection of fraudulent activities on stock market financial discussion boards

Ms. Pei Shyuan Lee
School of Computing, Mathematics & Digital Technology
The Manchester Metropolitan University, Chester Street,
Manchester, M1 5GD, UK
Email: Pei-Shyuan.Lee@mmu.ac.uk

Dr. Keeley Crockett
School of Computing, Mathematics & Digital Technology
The Manchester Metropolitan University, Chester Street,
Manchester, M1 5GD, UK
Email: K.Crockett@mmu.ac.uk

Dr. Majdi Owda
School of Computing, Mathematics & Digital Technology
The Manchester Metropolitan University, Chester Street,
Manchester, M1 5GD, UK
Email: M.Owda@mmu.ac.uk

Abstract*— Financial discussion boards (FDBs) have been widely used for a variety of financial knowledge exchange activities through the posting of comments. Popular public FDBs are prone to being used as a medium to spread false financial information due to larger audience groups. Although online forums are usually integrated with anti-spam tools such as Akismet, moderation of posted content heavily relies on manual tasks. Unfortunately, the daily comments volume received on popular FDBs realistically prevents human moderators to watch closely and moderate possibly fraudulent content, not to mention moderators are not usually assigned with such task. Due to the absence of useful tools, it is extremely time consuming and expensive to manually read and determine whether each comment is potentially fraudulent. This paper presents novel forward and backward analysis methodologies implemented in an Information Extraction (IE) prototype system named FDBs Miner (FDBM). The methodologies aim to detect potentially illegal Pump and Dump comments on FDBs with the integration of per-minute share prices in the detection process. This can possibly reduce false positives during the detection as it categorises the potentially illegal comments into different risk levels for investigation purposes. The proposed system extracts company's ticker symbols (i.e. unique symbol that represents and identifies each listed company on stock market), comments and share prices from FDBs based in the UK. The forward analysis methodology flags the potentially Pump and Dump comments using a predefined keywords template and labels the flagged comments with price hike thresholds. Subsequently, the backward analysis methodology employs a moving average technique to determine price abnormalities and backward analyse the flagged comments. First detection stage in forward analysis found 9.82% potentially illegal comments. It is unrealistic and unaffordable for human moderators or financial surveillance authorities to read these comments on a daily basis. Hence, by integrating share prices to perform backward analysis can categorise the flagged comments into different risk levels. It helps relevant authorities to prioritise and investigate into the higher risk flagged comments, which could potentially indicate a real Pump and Dump crime happening on FDBs when the system is being used in real time.

Keywords*— *Financial Discussion Boards; Financial Crimes; Pump and Dump; Text Mining; Information Extraction

I. INTRODUCTION

Internet has become the number one source for information for unlimited things. Unsurprisingly, this includes financial advice and investor sentiments. There are many online forums where likeminded people can hold conversations in the form of posted messages. Financial Discussion Boards (FDBs), also known as Financial Message Boards or Financial Forums allows investors to exchange knowledge, information, experience and opinions about the investment opportunities. There is a few popular share price based FDBs based in the UK which specifically allows investors to discuss share prices. These FDBs include the London South East[1], Interactive Investor (III)[2] and ADVFN[3].

Normally, forum content is moderated by human moderators when it is discovered or reported for breaching forum rules such as racism, sexism, hatred, foul language, third party advertisements and so on. Although online forums seem to be a useful source of information, not all information shared on the forums is accurate or truthful. Even anti-spam plugins such as Akismet[4] could only prevent spammers from registering or posting generic spam messages. There is little to no measurements taken by forum moderators or financial surveillance authorities to monitor and detect potential crimes on the FDBs, such as comments indicative of Pump and Dump (P&D).

[1] http://www.lse.co.uk
[2] http://www.iii.co.uk
[3] http://uk.advfn.com
[4] https://akismet.com

P&D can happen if an organised group of false investors decided to attack shares by buying and selling a specific share in a scheduled time frame and giving the market false statements about the share throughout the process. Textual comments such as "This is the right time let's start pumping this share" can reveal a hidden potential illegal activity of P&D on these FDBs. Novice investors can be easily deceived and make huge losses during the "dump" while the fraudsters take huge profits. Without a tool, manual monitoring and detection of potentially illegal activities on popular and active FDBs can cost time and money heavily, which is impracticable in the long run.

There has been research conducted around the area of share price based FDBs associated with P&D financial crimes [1, 2, 3, 4, 5, 6]. Research from recent years highlighted that the comments on FDBs were found manipulative and positively related to the market returns, volatility and trading volumes [7, 8, 9, 10, 11]. However, there is very little attempt [5, 6] made to build tools for monitoring and detection of potential financial crimes on share price based FDBs. Furthermore, other than the initial work presented in [12], none of the other existing research take share prices into account when designing a financial surveillance tool for detection of potentially illegal FDB comments.

FDBs contain semantically understandable artefacts (i.e. FDBs' artefacts that can be processed by computers) such as stock ticker symbols, date, time, prices, comment author usernames and comments. Information Extraction (IE) is defined as the process of extracting information automatically into a structured data format from an unstructured or semi-structured data sources [13]. Therefore, IE techniques are used in this research to extract and analyse these data. IE has been used in other areas such as accounting [14] and search engine [15]. However, other than the initial work described in [6] and [16], there is very little usage of IE techniques in FDBs' financial crimes related research.

Two novel methodologies, i.e. forward analysis and backward analysis, introduced in this paper are implemented in a prototype system named FDBs Miner (FDBM). The methodologies are used to detect potential P&D crimes on FDBs by flagging potentially illegal comments and reduce false positives (i.e. errors present in evaluation processes or scientific tests that are mistakenly found) during the detection process. FDBM could significantly support financial surveillance authorities to regulate by enabling real-time monitoring and alerting based on fraudulent risk levels.

In the forward analysis methodology, all the potentially illegal comments will first be highlighted and flagged. This is done by analysing the comments against the predefined P&D IE keywords template. Next, it matches and appends the price figure to the flagged comments which share the same or closest date and time based on same ticker symbol. Subsequently, the forward analyser takes each flagged comment's price as a base price and calculate ± 2 days' worth of prices to check if there is any price hike 5%, 10% and 15% more than the base price. Finally, it appends the price hike threshold labels to these flagged comments. By doing so, a relevant authority can pick the comments belong to any threshold depending on the severity for investigations. Although forward analysis has drastically reduced the number of comments needed to be read by relevant authorities, the amount of categorised flagged comments could still be somewhat large to read on a daily basis. Thus, a backward analysis methodology is designed to overcome this issue.

In the backward analysis methodology, simple moving average method is used to calculate and highlight the price hikes. Any price hikes that hit certain price hike thresholds will be matched backwards to the flagged comments found in the forward analysis stage. Such matches are done so that the already flagged comments can be further classified to reduce false positives and allow investigators to quickly examine on the higher and highest risked flagged comments before everything else.

Section II describes some examples of FDBs related financial crimes and reviews the background and usage of Information Extraction (IE) and Text Mining. Section III presents the architecture overview of the FDBs Miner (FDBM) prototype system and an overview of the FDBs dataset (FDB-DS). This followed by Section IV and V introducing the two novel methodologies (i.e. forward analysis and backward analysis) respectively and discussing the findings. Lastly, Section VI concludes the research and proposes some future work.

II. BACKGROUND

This section first provides a few related and significant examples of financial crimes on share price based FDBs, followed by the literature review related to IE and text mining which are the techniques used in this research for locating meaningful information, and collection and formation of datasets respectively. Lastly, Pump and Dump (P&D) and FDBs related literature review will also be presented.

A. Financial Crimes on Share Price Based FDBs

Generally, there are many P&D financial crimes happening which are actively investigated and dealt by the Security Exchange Commission (SEC) for many years. However, P&D crimes on FDBs are loosely monitored by FDB moderators and relevant authorities. There were several popular FDB related P&D financial crimes in early years, which are highlighted as follows:

- 15-year-old Jonathan Lebed was the first minor to involve in a stock market fraud in 2000 [3]. Lebed earned a total revenue of US$800,000 by pumping the share price through Yahoo! Finance Message Board over half a year and charged by Security Exchange Commission (SEC) [3, 4].
- In 2000, two were being charged for pumping the price of a share by 10,000% by posting on Raging Bull message board and then dumped millions of shares which the profit made was at least US$5 million [3].

- In addition in 2009, eight participants were charged by Security Exchange Commission (SEC)[5] for being involved in penny stock manipulation throughout the year of 2006 and 2007. These wrongdoers met each other through a popular penny stock message board.

Based on the above FDBs related P&D financial crimes, instead of investigating into the crimes after being committed – which is probably too late as the harm has been done - there is certainly a need to create methods and tools for detection of potentially illegal FDB comments in real time.

B. Information Extraction and Text Mining

This research makes use of Information Extraction (IE) and Text Mining. IE is being defined [17] as the process of extracting information automatically into a structured data format from an unstructured or semi-structured data sources. It was suggested [18] that there is a need for systems that extract information automatically from text data. IE is not Information Retrieval (IR) [19]. The difference between IE and IR is that IE extracts information that fits predefined templates or databases and then presents the information to the users, whereas IR finds data and present the information to the users. IE systems are knowledge-intensive as these systems extract only snippets of information that will fit predefined templates (fixed format) which represent useful and relevant information about the domain then display to the user.

IE is divided into two fundamental classes i.e. Knowledge Engineering (KE) approach and automatic training approach. KE approach is also called as the rule-based approach since it requires rules to be developed by the human expertise. Rule-based approach is usually ignored in the research community, but it is mostly favourable in the commercial market even by the large vendors such as IBM (for text analysis systems) and Microsoft (enterprise search platform) [20]. Rule-based IE systems are easy to maintain and comprehend as well as errors being traced and fixed easily. On the other hand, although automatic training approach, also known as machine learning approach, requires less manual efforts, the approach requires pre-labelled data and retraining for adaptation [20]. This paper focuses on IE implementation since it is designed to support the financial market surveillance authorities.

Text mining was described [21] as the process to extract useful information from unformatted textual data or natural language text into a form of meaningful knowledge for processing. According to [22], the research shows that there was a significant amount of users on Twitter (32%) and Facebook (20%) were actively seeking or sharing advice about their favourite products at least once a week. This means the likelihood of getting deceptive information is also significant. Similarly, on popular share price based FDBs that receive a significant amount of comments in each day, novice investors who seek investment advice could also be deceived easily. Also a text mining based study was conducted [23] on Twitter dataset and its relationship to be able to predict stock prices. In addition to stock price trends were also being successfully forecasted via press releases using text mining techniques [24].

[5] http://www.sec.gov/litigation/litreleases/2009/lr21053.htm

In this paper, text mining is used alongside IE rule-based technique to extract and analyse FDB artefacts such as comments, prices and stock ticker symbols.

C. Pump and Dump and Share Price Based FDBs

Traditionally, Pump and Dump (P&D) happens mostly through word of mouth. But with the existence of the Internet, it becomes so common that the fraudsters commit crimes through various channels such as emails, discussion boards and social media.

As spam emails is one of the older tactics, regulators like Securities and Exchange Commission (SEC) has been actively taking actions against P&D spam campaign fraudsters. Email spam filters are also constantly being improved by Internet services such as Google and Symantec. In a research [25], a total of 1,299 suspicious stock recommendation emails was obtained. It involved 221 stocks recommended in 252 advertising campaigns. An event study and a sentiment analysis have been conducted on whether P&D involving the internet is still an issue in today's world. Unsurprisingly, the research empirically proved that the internet still plays a major role in enhancing this type financial crime. Due to the limitations in spam emails, newer tactics such as social media and discussion boards were adopted mainly because these channels allow more freedom of speech. Other researchers [7, 8, 9, 10, 11] have found the relation between FDB comments and market performance. FDB comments can be manipulative and affect the share prices.

In [5], the authors introduced a novel classification technique for a classifier training in order to automate moderation tasks on online discussion sites (ODSs). A partially labelled corpus is used for the training purpose and then attempt to moderate the inappropriate content on ODSs using the technique. The authors implemented and tested the technique on a corpus of comments posted on a popular Australian FDB named HotCopper[6]. The results indicated that the classification technique is helpful and can be used to decrease the number of comments that need to be moderated by human moderators. However, this system is not yet a fully automated moderation system due to the use of partially labelled corpus. According to the authors, the misclassification errors remain too significant. Besides, the research takes only comments into account and no prices involved during the classification of comments.

A system named Financial Discussions Detection System (FDDS), an initial work to this research, was proposed by the authors in [6] to flag potentially illegal comments made on FDBs. The system allows users to create and modify predefined templates (i.e. lists of potentially illegal keywords that commenters may or frequently use on FDBs), download comments from FDBs and matches the downloaded comments against the potentially illegal keywords created in earlier steps. By looking only at the comments during the detection processes appear to be insufficient in terms of accuracy. Thus, this paper introduces the novel methodologies in attempt to reduce false positives by integrating share prices in the detection process.

[6] https://hotcopper.com.au

The authors in [11] examined whether the messages posted on the largest stock message board in Australia, HotCopper, has an impact on the Australian Stock Exchange (ASX) market. Results show that the FDB messages have impacts on the small capitalisation stocks but not affecting the large stocks.

In [26], the authors introduced a software prototype (FMS-DSS) to support decision making in financial market surveillance. FMS-DSS consists of three components i.e. data, models and user interface. The system collects both unstructured and structured data of the selected listed companies. The models take into account of attributes such as market segment, market capitalisation, trading volume, age of company and so on. Subsequently, attribute scales ranging from very low to very high were defined by the regulatory authority members. The scales were then used for aggregation to determine whether there is suspicious activity happening.

In attempt to resolve what was missing in existing research, share prices are taken into account when flagging potentially illegal comments, accompanied by two key novel built-in methodologies (namely the forward analysis and the backward analysis) for resolving false positives during the comments flagging process.

III. Architecture Overview

This section presents the FDBM architecture which consists of several key components. These key components are the data crawler, data transformer, FDB dataset (FDB-DS), IE keyword template, forward analyser and the backward analyser. Fundamentally, FDBM collects data, transform unstructured data into structured data format and analyse the data using both forward and backward analysers. The forward analyser and backward analyser components are used within the novel methodologies introduced in this paper attempt to resolve false positives during the process of detection of potentially illegal comments.

A. Overview

Figure 1 provides an overview of the FDBM architecture of the prototype system.

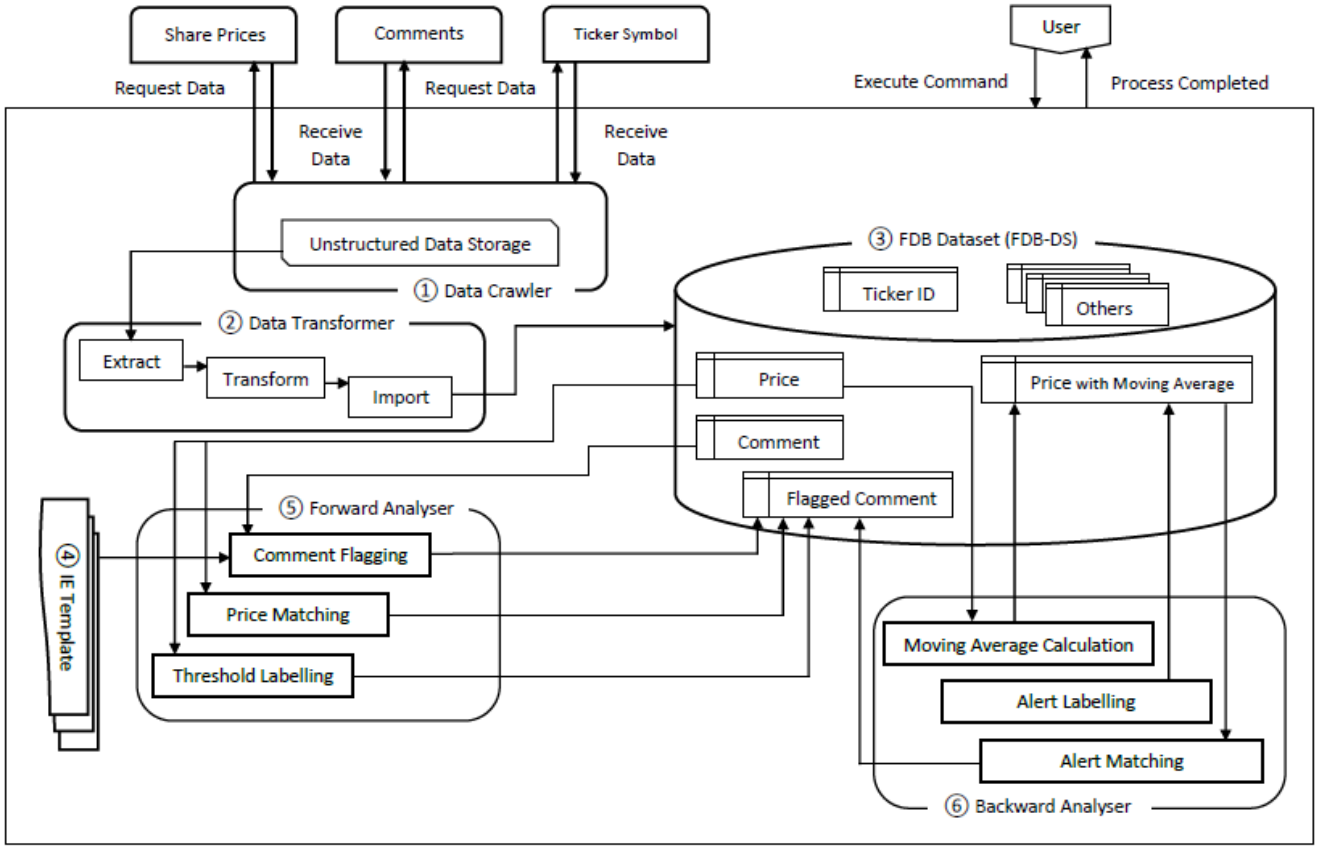

Fig. 1. Architecture Overview Diagram.

Each component in the architecture diagram is described as follows:

1. **Data Crawler -** The data crawler is responsible for automatically collecting unstructured data from the three FDBs (i.e. LSE, III and ADVFN) at different time intervals for 12 weeks (from 23rd September 2014 to 22nd December 2014). These unstructured and semi-structure data consist of 941 ticker symbols that were listed on London Stock Exchange (LSE), FTSE100 and FTSE AIM All-Share, 1-minute bar price figures for all the 941 companies and all the available FDB comments belong to the 941

companies. As an effort for potential future work, director deals data and broker ratings data were also collected. Table 1 in Section B summarises the total sum of collected data.

2. **Data Transformer** - Once the data collection is done by the data crawler, the data transformer extracts and converts the collected unstructured data in various formats such as HTML, CSV and XML into structured data.

3. **FDB Dataset (FDB-DS)** – After the collected data is being processed by the data transformer, the structured data such as price figures, comments, comment author usernames, date and time of comments and prices are stored in the FDB-DS accordingly. For example, the ticker symbols are parsed into `ticker` table, price data are parsed into `price` table and comment data are parsed into `comment` table. The FDB-DS is also responsible to store additional data produced from research analysis.

be easily modified whenever needed. The IE keyword template consists of a series of keywords and phrases that were thoroughly researched [2, 27, 28, 29] and has been validated by experts in the relevant field. The IE keyword template will be used by the forward and backward analysers for the comments flagging process. Section C shows a sample list of the keywords and phrases.

5. **Forward Analyser –** The forward analyser matches the Pump and Dump IE keyword template against the comments in order to flag potentially illegal FDB comments. Followed by matching the prices to the flagged comments, calculating and labelling price thresholds. The novel methodology used in this component is further discussed in Section IV.

6. **Backward Analyser** – Backward analyser performs the calculation and labelling of price hikes using a price moving average technique i.e. simple moving average (SMA). This calculation is applied against a

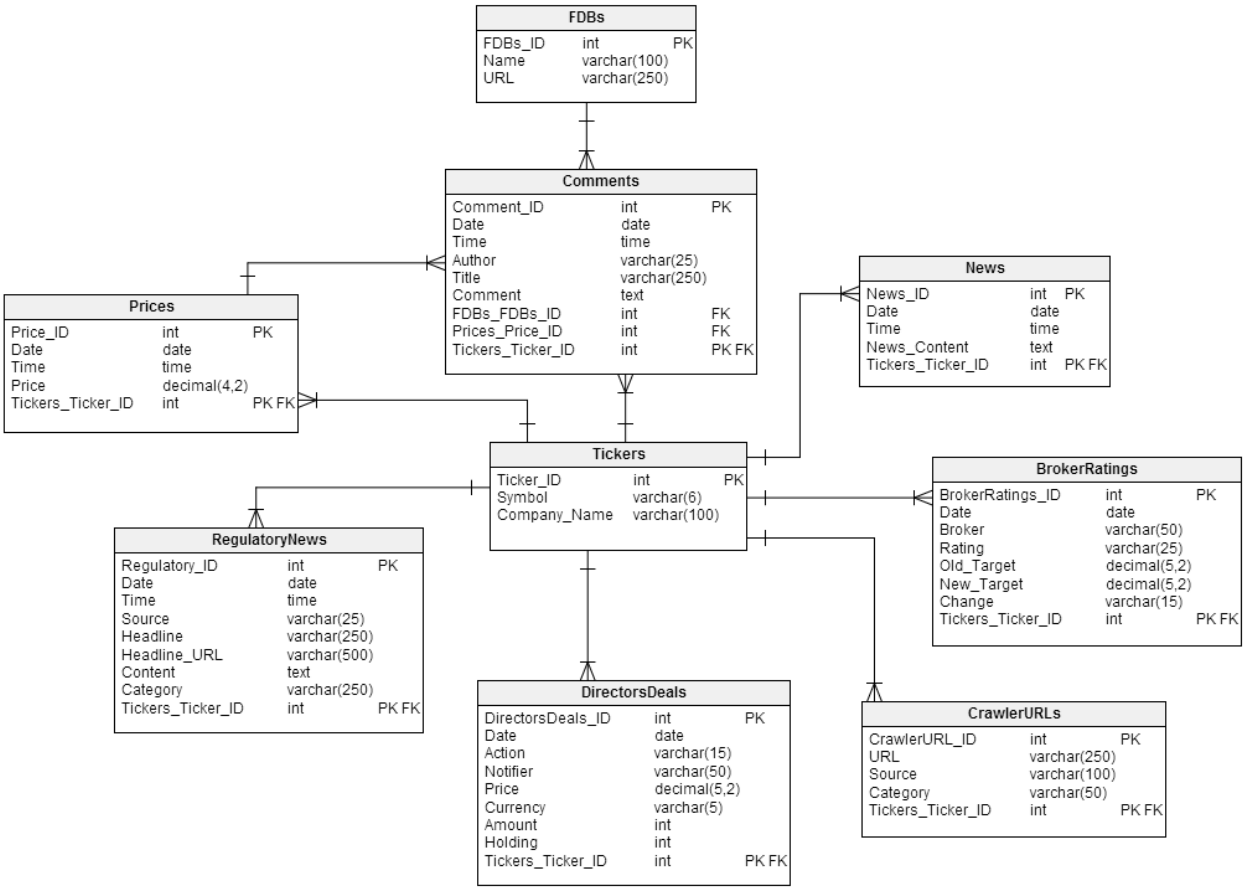

Fig. 2. FDB Dataset Structure.

4. **IE Templates –** The Pump and Dump IE keyword template has been created and saved locally in the prototype system in a text (TXT) file format. It can

total of 29 million price figures which belong to 941 companies. Subsequently, price hike SMA alerts will

be matched back towards the initially flagged comments in forward analysis process. This methodology is further elaborated in Section V.

B. Dataset Acquisition

Table 1 provides an overview of the FDB dataset (FDB-DS) in this research. These data were collected between 23rd September 2014 to 22nd December 2014.

TABLE 1. TOTAL NUMBER OF ARTEFACT RECORDS (FDB-DS)

Artefacts	Total number of records
Ticker Symbols	941
Comments	507,970
Prices	28,980,465
Director Deals	11,456
Broker Ratings	6,469

As mentioned in earlier section, these 941 ticker symbols were collected from two of the LSE's indices i.e. 100 ticker symbols from FTSE100 and 841 ticker symbols from FTSE AIM All-Share. The comments, which belong to all these ticker symbols, made within the 12 weeks were collected from both LSE and III. As for prices, these are 12 weeks' worth of 1-minute bar share prices belong to all the 941 ticker symbols. Director deals and broker ratings related to all the ticker symbols were also collected for potential future work. The following is an overview of the FDB-DS structure.

C. IE Template

Pump & Dump (P&D) IE keyword template is populated by obtaining the keywords from the P&D comments demonstrated in existing research [6, 27, 28, 29]. The following is a sample list of the keywords and phrases:

- Pump dump
- Once in a lifetime
- Pump the price
- Keep ramping
- Buy now
- Good future
- Invested so heavily
- It will fly
- Sell now
- This is the chance
- Price will go up
- Buy as quickly as possible
- Get out while you can

IV. FORWARD ANALYSIS METHODOLOGY

This section introduces the novel forward analysis methodology. The aim of this methodology is to flag and filter the potentially illegal P&D comments using P&D keyword template with the integration of the share prices in the analysis process. This will categorise the flagged comments into different risk levels and allows relevant authorities to investigate into the flagged comments more realistically in terms of time and efforts.

The forward analysis methodology in this section will test the following hypothesis:

H_{0a}: Pump and Dump activity from FDBs can be filtered using template based IE and their correlation with price movements.

H_{1a}: Pump and Dump activity from FDBs cannot be filtered using template based IE and their correlation with price movements.

As shown in the architecture diagram in Figure 1, the forward analysis component contains several functions. These functions (i.e. comments flagging, price matching, threshold calculation and threshold labelling) that are part of the forward analysis methodology which will be discussed below.

A. Methodology

The following describes the steps taken in this methodology to flag potentially illegal comments:

1) Comments Flagging

i. Firstly, the forward analyser matches all the available keywords and phrases from the Pump and Dump IE keyword template against all the 507,970 comments which were stored in FDB dataset (FDB-DS).

ii. The flagged comments which deemed potentially illegal are imported into FDB-DS as a new database table named `flaggedcomment`.

2) Price and Comments Matching

i. Once `flaggedcomment` has been populated, the forward analyser appends the price to each flagged comment by matching the ticker symbol and the exact or nearest date and time. This step is done to ensure a "base price" is set for each flagged comment. The "base price" will be used for threshold labelling in next step. Due to the extremely large 12 weeks' worth of price data belongs to 941 companies, the process of setting a "base price" takes up to a week to complete.

3) Comments Threshold Labelling

i. After having all the "base price" set for each flagged comment in the previous step, the forward analyser labels each flagged comment with thresholds. Due to the large data set, the threshold labelling process takes up to five days to complete all threshold calculations. To determine whether a flagged comment's base price exceeds any thresholds (i.e. various levels of spikes in prices), the forward analyser calculates all the ± 2 days' per-minute prices against the "base price" of each flagged comment.

ii. When there is a trigger, a flagged comment will be labelled accordingly. The threshold labelling rules are as follows:

- Flagged comments that have no price figure (due to empty price figures collected from ADVFN) is labelled as "N" (Null).
- If any of the ± 2 days prices calculated against the "base price" indicates a 5% price hike the comment is labelled as "Y" (Yellow).
- If any of the ± 2 days prices calculated against the "base price" indicates a 10% price hike the comment is labelled as "A" (Amber).
- If any of the ± 2 days prices calculated against the "base price" indicates a 15% price hike the comment is labelled as "R" (Red).
- Flagged comments that do not trigger any thresholds are labelled as "C".

B. Forward Analysis Methodology Results

By matching the keywords and phrases from P&D IE keyword template against all the 507,970 comments, a total number of 49,858 comments were flagged as potentially illegal comments (as shown in Table 2). These flagged comments took up 9.82% of the total comments.

TABLE II. TOTAL NUMBER OF FLAGGED COMMENTS

Comments	Total	Percentage
Flagged	49,858	9.82%
Non-flagged	458,112	90.18%
Grand Total	507,970	100%

Out of all the 49,858 flagged comments, 3,613 (7.25%) of the flagged comments triggered the "R" 15% price hike threshold, 2,555 (5.12%) flagged comments triggered the "A" 10% price hike threshold and 5,197 (10.42%) flagged comments triggered the "Y" 5% price hike threshold. 37,895 (76.01%) flagged comments labelled as "C" did not trigger any price thresholds. The total number of flagged comments that triggered the thresholds is summarised in Table 3 and visualised in Figure 3.

TABLE III. TOTAL NUMBER OF FLAGGED COMMENTS IN EACH PRICE HIKE THRESHOLD

Threshold	Total	Percentage
C (<5%)	37,895	76.01%
Y (5%)	5,197	10.42%
A (10%)	2,555	5.12%
R (15%)	3,613	7.25%
Null	598	1.2%
Grand Total	49,858	100%

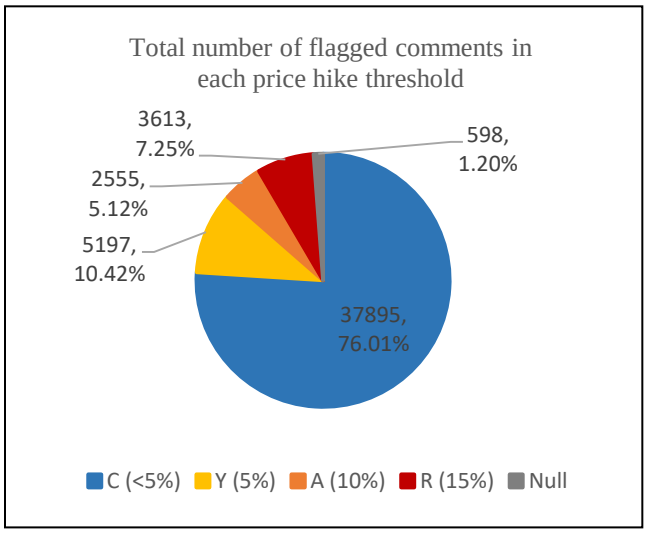

Fig. 3. Total number and percentage of each threshold.

The results show the possibility to filter comments that may be indicative of Pump and Dump activities by using template based IE and the correlation with price movements. For 12 weeks' worth of 941 companies' share prices data, the forward analyser took approximately seven days to completely calculate all the price thresholds and labelling the flagged comments. The length of time taken in this process heavily relied on the computer machine power and the efficiency of the programming in FDBM. In this research, the server machine used is a quad core CPU (2.50GHz Intel(R) Xeon(R) CPU E5-2680 v3). Although the forward analysis process takes a long time to process, this is due to the massive amount of data being processed altogether in this research. In real world scenario, this methodology can significantly help relevant authorities to narrow down and focus on the potentially illegal comments with higher risks. Therefore, the hypothesis for this section is met.

V. BACKWARD ANALYSIS METHODOLOGY

As an enhancement to the forward analysis process, the novel backward analysis process will test whether simple moving average (SMA) technique can be used to reduce false positives in the comments flagging process by highlighting abnormalities in the share prices and backward classify the flagged comments.

The backward analysis methodology in this section will test the following hypothesis:

H_{0b}: Backward analysis can be performed by matching abnormal stock prices with the flagged comments to further classify flagged comments to reduce false positive.

H_{1b}: Backward analysis cannot be performed by matching abnormal stock prices with the flagged comments to further classify flagged comments to reduce false positive.

Moving average is one of the technical analysis methods that is often being used by financial analysts to predict the future price patterns, learning stocks' behaviour and trends by studying historical price data. The most basic moving average technique being used by financial analysts is SMA. Some research even used such moving average techniques to predict the rate of traffic congestions and road accidents [30]. However, it appears that there was no attempt to integrate moving average technique in the detection process of potential FDB crimes in the past.

The backward analysis attempts to use SMA to test if it can be of helpful to detect flagged comments while reducing false positives. SMA technique is integrated and applied to the share prices before performing backward analysis. Moving average technique is used in backward analysis because it can calculate and highlight whether a price figure exceeds a certain threshold. The following section discusses the methodology to perform backward analysis.

A. Methodology

The following describes the steps taken to produce results for analysis:

1) *Moving Average Calculation*
   i. Firstly, decide time periods use for this experiment i.e. 1 day, 3 days and 5 days.
   ii. Next, calculate the Simple Moving Average (SMA) using its formula as below and record calculation results in database:

$$SMA = \frac{p_1 + p_2 + \cdots + p_n}{n}$$

2) *Alert Labelling*
   i. Apply 5%, 10% and 15% thresholds calculation based on the calculated SMA figures and save in database table. Table III shows an example of the threshold calculations, assuming the SMA is $15.4:

TABLE III. SMA THRESHOLD CALCULATION EXAMPLE

Threshold	SMA Threshold Price
5%	$15.4 * 1.05 = $16.17
10%	$15.4 * 1.10 = $16.94
15%	$15.4 * 1.15 = $17.71

   ii. Once the SMA figures and threshold figures above SMA are obtained, check each original price against the calculated threshold figures. If an original price exceeds the calculated threshold figure, label these threshold alerts accordingly (i.e. 5%, 10% or 15%). The alert labelling rules are as follows:

   - Label as "5%" - if the original price figure of a particular date and time is between 5% and 10% higher than the SMA price figure.
   - Label as "10%" - if the original price figure of a particular date and time is between 10% and 15% higher than the SMA price figure.
   - Label as "15%" - if the original price figure of a particular date and time is 15% and above the SMA price figure.

3) *Alert Matching*
   i. Next, the backward analyser appends the price alerts back to the `flaggedcomment` table by matching the ticker symbol and the exact or nearest date and time between both `price` and `flaggedcomment` tables.

B. Backwards Analysis Methodology Results

Table IV shows the total number of flagged comments that matched 5% threshold from both forward and backward analysis for the 1 day, 3 days and 5 days' time period. Out of 49,858 flagged comments there are 228 flagged comments from the 1 day time period experiment labelled with Y (5% threshold from forward analysis) which are also labelled with 5% threshold from backward analysis. Next, there are 306 flagged comments from the 3 days' time period labelled with Y (5% threshold from forward analysis) and 5% threshold from backward analysis. Lastly, there are 274 flagged comments from the 5 days' time period labelled with Y (5% threshold from forward analysis) and 5% threshold from backward analysis.

TABLE IV. TOTAL NUMBER OF FLAGGED COMMENTS THAT MATCHED 5% THRESHOLD FROM BOTH FORWARD AND BACKWARD ANALYSIS

	5% 1D	5% 3D	5% 5D
C (<5%)	518	1039	1300
Y (5%)	228	306	274
A (10%)	89	259	183
R (15%)	154	126	84

Table V shows the total number of flagged comments that matched 10% threshold from both forward and backward analysis for the 1 day, 3 days and 5 days' time period. Out of 49,858 flagged comments there are 40 flagged comments from the 1 day time period experiment labelled with A (10% threshold from forward analysis) which are also labelled with 10% threshold from backward analysis. Next, followed by 49 flagged comments from the 3 days' period labelled with A (10% threshold from forward analysis) and 10% threshold from backward analysis. Lastly, there are 64 flagged comments from the 5 days' period labelled with A (10% threshold from forward analysis) and 10% threshold from backward analysis.

TABLE V. TOTAL NUMBER OF FLAGGED COMMENTS THAT MATCHED 10% THRESHOLD FROM BOTH FORWARD AND BACKWARD ANALYSIS

	10% 1D	10% 3D	10% 5D
C (<5%)	204	291	366
Y (5%)	99	62	100
A (10%)	40	49	64
R (15%)	79	85	97

Table VI shows the total number of flagged comments that matched 15% threshold from both forward and backward analysis for the 1 day, 3 days and 5 days' period. Out of 49,858 flagged comments there are 199 flagged comments from the 1 day time period experiment labelled with R (15% threshold from forward analysis) which are also labelled with 15% threshold from backward analysis. There are 408 flagged comments from the 3 days' time period labelled with R (15% threshold from forward analysis) and 15% threshold from backward analysis. Lastly, there are 500 flagged comments from the 5 days' time period labelled with R (15% threshold from forward analysis) and 15% threshold from backward analysis.

TABLE VI. TOTAL NUMBER OF FLAGGED COMMENTS THAT MATCHED 15% THRESHOLD FROM BOTH FORWARD AND BACKWARD ANALYSIS

	15% 1D	15% 3D	15% 5D
C (<5%)	242	356	395
Y (5%)	74	127	146
A (10%)	42	65	94
R (15%)	199	408	500

The results in Table IV, V and VI show it is possible to perform backward analysis by matching the abnormal stock prices backwards to the flagged comments to resolve false positives.

Take ticker symbol "BOX" as an example, there are 50 comments belong to this stock flagged as "R (15%)" threshold in the forward analysis process. Subsequently, some of these comments are flagged with SMA 15% threshold alert in the backward analysis process. This indicates that there are very high chance of potentially illegal activities going on during ± 2 days' time of the comments made. A further look at these flagged comments can confirm a highly potential P&D crime. One comment suggests that P&D has indeed happened which pumped the price up and then dumped. Another comment shows that there is still an attempt to pump up the price after the P&D event. Author "ne14t" has a series of BOX comments showing that he/she could possibly involve in a P&D crime.

Date/time: 06/10/2014 14:42:38
Author: bigwod
Comment: slow build up is what i wanted had some fools ramp it up and it was gone now its back

Date/time: 07/10/2014 09:02:19
Author: ne14t
Comment: buys now showing the correct colour!

As an enhancement to forward analysis methodology, backward analysis aims to resolve false positives and reduce the need of a lot of manpower and time to read through initially flagged comments. The time taken in both forward and backward analysis process in this research is long; however, this is only due to the significant amount of data being processed and analysed altogether. If the prototype system and both methodologies are applied in real time in real world scenarios, it can significantly reduce the time, effort and cost of monitoring and detecting P&D crimes on FDBs. Therefore, this concluded that the hypothesis is met.

VI. CONCLUSION AND FUTURE WORK

This paper has introduced two novel methodologies for detecting potentially illegal activities on share price based FDBs by looking not only at the comments but also the per minute share prices. IE techniques were used to collect FDB artefacts such as ticker symbol, comments and prices which made the forward analysis possible to be conducted in this research. A total of 49,858 comments were flagged when matching against the P&D IE keyword template. In average, this is 4,154 flagged comments per week or 593 flagged comments a day. More importantly, these comments belong to only 941 listed companies, not the entire stock market in the UK. In order to perform a more realistic investigation into such financial crime on all the FDBs and for all listed companies in the UK on a daily basis, the forward and backward analysis methodologies integrate share prices in the analysis process. This makes it possible for the relevant authorities to prioritise on investigating the flagged comments that have higher risks. The methodologies implemented in FDBM can significantly reduce the time and efforts needed by the relevant authorities to investigate P&D crime on FDBs in real time. This research considers integrating Semantic Textual Similarity (STS) technique into our overall methodology as part of the near future work.

VII. REFERENCES

[1] Leinweber, D.J., & Madhavan, A.N., "Three Hundred Years of Stock Market Manipulations," Journal of Investing, p. 7–16, 2001.

[2] Campbell, J.A., "In and Out Scream and Shout: An Internet Conversation about Stock Price Manipulation," Proceedings of the 34th Hawaii International Conference on System Sciences, p. 1–10, 2001.

[3] Riem, A., "Cybercrimes of the 21st Century: Crimes against the individual — Part 1. Computer Fraud & Security," 6, 13–17, 2001.

[4] Cybenko, G., Giani, A., & Thompson, P., "Cognitive Hacking: A Battle for the Mind," 2002.

[5] Delort, J. Y., Arunasalam, B., & Paris, C., "Automatic Moderation of Online Discussion Sites," International Journal of Electronic Commerce, 15(3), p. 9–30, 2011.

[6] Knott, E., & Owda, M., "The detection of potentially illegal activity on financial discussion boards using information extraction," 2nd International Conference on Cybercrime, Security and Digital Forensics, London, UK, 2012.

[7] Antweiler, W., & Frank, M. Z., "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," The Journal of Finance, 59(3), p. 1259–1294, 2004.

[8] Cook, D. O., & Lu, X., "Noise, Information, and Rumors: Internet Boards Messages Affect Stock Returns," University of Alabama, 2009.

[9] Delort, J. Y., Arunasalam, B., & Leung, H., "The Impact of Manipulation in Internet Stock Message Boards," International Journal of Banking and Finance, 8(4), p. 1–18, 2011.

[10] Bettman, J., Hallett, A., & Sault, S., "Rumortrage: Can Investors Profit on Takeover Rumors on Internet Stock Message Boards?", 2011.

[11] Leung, H., and Ton, T., "The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks," Journal of Banking & Finance, p. 37–55, 2015.

[12] Lee, P. S., Owda, M., & Crockett, K., "The detection of fraud activities on the stock market through forward analysis methodology of financial discussion boards," Future of Information and Communications Conference, Singapore, 2018.

[13] Cowie, J., & Lehnert, W., "Information Extraction," Communications of the ACM, 39(1), p. 80–91, 1996.

[14] Seo, K., Choi, J., & Choi, Y., "Research about Extracting and Analyzing Accounting Data of Company to Detect Financial Fraud. Intelligence and Security Informatics, p. 200–202, 2009.

[15] Limanto et al, "An Information Extraction Engine for Web Discussion Forums," Nanyang Technological University, Singapore. ACM 1-59593-051-5/05/0005, May 2005.

[16] Owda, M., Lee, P. S., Crockett, K., "Financial Discussion Boards Irregularities Detection System (FDBs-IDS) using Information Extraction," Intelligent Systems Conference 2017, London, UK, 2017.

[17] Masterson, D., & Kushmerick, N., "Information Extraction from Multi-Document Threads," 2003.

[18] Soderland, S., "Learning Information Extraction Rules for Semi-structured and Free Text," 1999.

[19] Cunningham H., "Information Extraction, Automatic," in Keith Brown, (Editor-in-Chief) Encyclopedia of Language & Linguistics, Second Edition, 5, p. 665-677, 2006.

[20] Chiticariu, L., Li, Y., & Reiss, R. F., "Rule-based Information Extraction is Dead! Long Live Rule-based Information Extraction Systems!," Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, p. 827–832, Seattle, Washington, USA, 2013.

[21] Kaiser, C., & Bodendorf, F., "Mining consumer dialog in online forums," Internet Research, 22(3), p. 275-297, 2012.

[22] Westerman, D., Spence, P. R., & Van Der Heide, B., "Social Media as Information Source: Recency of Updates and Credibility of Information," Journal of Computer-Mediated Communication, 19, p. 171-183, 2014

[23] Wolfram, M. S. A., "Modelling the Stock Market using Twitter," 2010.

[24] Mittermayer, M., "Forecasting Intraday Stock Price Trends with Text Mining Techniques," in Hawai'i International Conference on System Sciences, 2014.

[25] Siering, M., "All Pump, No Dump? The Impact of Internet Deception on Stock Markets," ECIS 2013 Completed Research, 115, 2013.

[26] Alić, I., "Supporting Financial Market Surveillance: An IT Artifact Evaluation, BLED 2015 Proceedings, Paper 36, 2015.

[27] Felton, J., & Kim, J., "Warnings from the Enron Message Board," Journal of Investing, 11(3), p. 29-52, 2002.

[28] Campbell, J.A. & Cecez-Kecmanovic, D., "Communicative practices in an online financial forum during abnormal stock market behavior. Information and Management, 48, p. 37-52, 2011.

[29] Sabherwal, S., Sarkar, S.K., & Zhang, Y., "Do Internet Stock Message Boards Influence Trading? Evidence from Heavily Discussed Stocks with No Fundamental News," Journal of Business Finance & Accounting, 38(9) & (10), p. 1209–1237, 2011.

[30] Raiyn, J., and Toledo, T., "Real-Time Road Traffic Anomaly Detection," Journal of Transportation Technologies, 4(3), p. 256-266, 2014.