Financial Discussion Boards Irregularities Detection System (FDBs-IDS) using Information Extraction

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

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

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

***Abstract*— The current growth and the use technology in global stock markets has created unprecedented opportunities for the individuals and businesses to access capital and grow and diversify their portfolios. Individuals, nowadays can decide to invest and act in few minutes if not in few seconds. This growth has led to a corresponding growth in the amount of fraud and misconduct seen in the stock markets through the use of technology. The internet is often used as a real time platform for illegal financial activity such as illegal activities on Financial Discussion Boards (FDBs). Managing and monitoring FDBs in real time is a complex and time consuming task; given the volume of data produced and the fact that some of the data is unstructured. This paper presents a novel Financial Discussion Boards Irregularities Detection System (FDBs-IDS) for FDBs which can highlight irregularities or potentially unlawful practices on FDBs. For example comments that might suggest a pump and dump activity is happening. The proposed system extracts information from FDBs, where templates hosting scenarios of known illegal activities are used to detect any potential misdemeanors. Analysis conducted on a single day trading, found that of the 3000 comments extracted from FDBs, 0.2% of these comments were deemed suspicious and required further investigation of a discussion board moderator. The man-power required to perform this task manually over the course of a year could be excessive and unaffordable. This research highlights the importance and the need of an automated crime detection system on FDBs such as FDBs-IDS which could be used and thus tackle potential criminal activities on the internet.**

Keywords — Information Extraction, Financial Discussion Boards, Fraud Detection, Financial Fraud Online, Crime Prevention, Text Mining, Financial Discussion Boards Mining and Web Mining.

I. INTRODUCTION

Financial Discussion Boards (FDBs) on the internet grant users commentary and subsequent discussion opportunities centering on shares, stocks, common funds and business in general [1, 2, 3]. These forums are not moderated by external third parties and loosely self-moderated via the forums' users themselves and administrators [3]; whether it will be a user reporting a comment as inappropriate, for instance. These forms of unmoderated communication are open to abuse and could play a significant part in the aiding and abetting of financial misconduct [4, 5, 6, 7, 8, 9]. Financial misconduct and crimes such as Pump and Dump and Insider Information can be found in these FDBs [6, 10]. Comments such as "This is the right time let's start pumping this share" can reveal a hidden potential illegal activity of Pump and Dump. Potentially illegal comments on FDBs were found to be manipulative and positively related to the market returns, volatility and trading volumes [6]. Artificial Intelligence (AI) has been widely employed in many financial fraud applications such as credit card fraud detection [11], and stock price forecasting [12] yet limited research has been conducted on stock market irregularities detection from the FDBs. FDBs have a number of unique features, named entities and process-able artifacts of the domain in which makes the data process-able by computers such as unique stock ticker name.

Information Extraction (IE) is a cascade of sequential steps, at each of which the system will add a structure and often lose information [13]. IE has been used in various fields in recent years often to extract key facts and for reasoning based on specific representation, notably: Text Mining [14] and bioinformatics [15]. It appears that very little research has been conducted with specific reference to IE within FDBs for the analysis of potentially illegal activity other than initial work reported in [10].

The solution presented in this paper could significantly impact the way FDBs are regulated in the future. The paper will outline why a proposed system is needed and how it has the potential to automatically tackle fraudulent activity born out of seemingly innocuous exchanges on FDBs. This paper proposes a new real time monitoring system of FDBs for

irregularities and potentially illegal practices detection called the Financial Discussion Boards Irregularities Detection System (FDBs-IDS).The key contribution of this work is introducing a methodology and a tool for automatically highlighting potential irregular activities on FDBs in real time in which it will reduce significantly the time needed for fraud investigators to reveal fraudulent activities on FDBs.

Section 2 introduces the concept of stock market fraud. Section 3 introduces and critically reviews financial discussion boards and their relationship with stock market fraud. Section 4 will introduce Information Extraction methods and critically review their suitability for FDBs-IDS. Section 5 will outline the FDBs-IDS system architecture. Section 6 will introduce the implementation. Section 7 will introduce the results and Section 8 will conclude the research outputs.

II. Stock Markets Fraud

The current growth and the use technology in global stock markets have created unprecedented opportunities for businesses to access capital and investors to grow and diversify their portfolios [16]. Individuals nowadays can invest through a number of channels such as their individual brokerage accounts, savings plans, or retirement accounts. This growth has led to a corresponding growth in the amount of fraud and misconduct seen in the stock markets [16]. In the United States, according to the Federal Bureau of Investigation (FBI) statistics shown in figure 1. The security and commodity based fraud pending cases in which stock mark fraud falls within this category; are on the increase year by year. This provides a clear justification for the need for innovative solutions to combat such illegal activities.

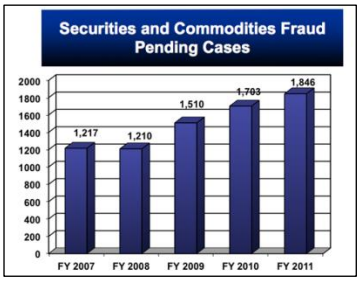

Fig. 1. Securities Fraud Pending Cases 2007-2011 [16]

III. Financial Discussions Boards

Approximately 5000 [1] comments are posted daily on only one of the active FDBs in the UK. To moderate such volume manually is extremely difficult. FDBs are usually moderated by the website administrators or volunteers in which they are not capable to perform real time monitoring of contents being posted. There have been a number criminal cases which have involved the usage of FDBs in order to make illegal financial gains such as Jonathan Lebed I, whom was involved in a stock market fraud in 2000; Lebed earned a total revenue of $800,000 US dollar by pumping stock prices through Yahoo Finance Message Board over half a year [17, 18]. In 2009, eight participants were charged by the Security Exchange Commission (SEC) for being involved in penny stock manipulation . These wrongdoers met each other through InvestorsHub.com, a popular penny stock message board, and carried out the Pump and Dump scheme throughout the year of 2006 and 2007. This case demonstrations one good reason for the wrongdoers to take public forums for granted to organise financial crimes by recruiting or meeting people who are willing to become part of the actual crime.

IV. Information Extraction (IE)

IE is a type of information retrieval where a user can define specific information to be extracted from documents (i.e. using a set of criteria, usually text, as opposed to images and videos). It can be associated with any method whose purpose is to extract information from documents and/or web pages. Chelba and Mahajan [19] defined IE as a text filtering and template filling process, segments of text are to be filled into a specific number of slots which forms a template or frame.

IE has two basic approaches; knowledge engineering and automatic training. First, the knowledge engineering approach is based on having a knowledge engineer who develops rules and knowledge that have the ability to solve problems in the real world for a specific domain. Appelt and Israel [13] believe that the knowledge engineering based approach is most effective when resources such as lexicons and rule writers are available. Secondly, the automatic training approach does not require a human to write rules for the IE system, instead, it only requires someone who knows the domain well, and then the task is to annotate a corpus of texts for the information being extracted. IE will play an important role in developing the financial discussion boards irregularities detection system (FDBs-IDS) described in the following section. In addition FDBs offer a number of unique artifacts in which an automated system could process and reason based on their availability and associated values; these artifacts such as stock unique name, stock price and comments.

V. Financial Discussion Boards Irregularities Detection System(FDBs-IDS) Architecture

This section will present the novel real time FDBs-IDS system architecture shown in figure 2 and will provide a detailed explanations of all components. In general, the FDBs-IDS system will be initialized by the user sending a query to retrieve potential irregular activity. The query processor then utilises the extractor to retrieve comments on FDBs in real time and stores them into a database. Then query processor begins analyzing the comments, it uses a lexicon in order to populate predefined IE templates for potentially illegal activities or irregular activities. Once a template has been

filled by the query processor using the database, lexicon and the IE templates, a response is sent to the user for analysis. More detailed explanations of the components used in the architecture shown in figure 2 are explained in the following subsections.

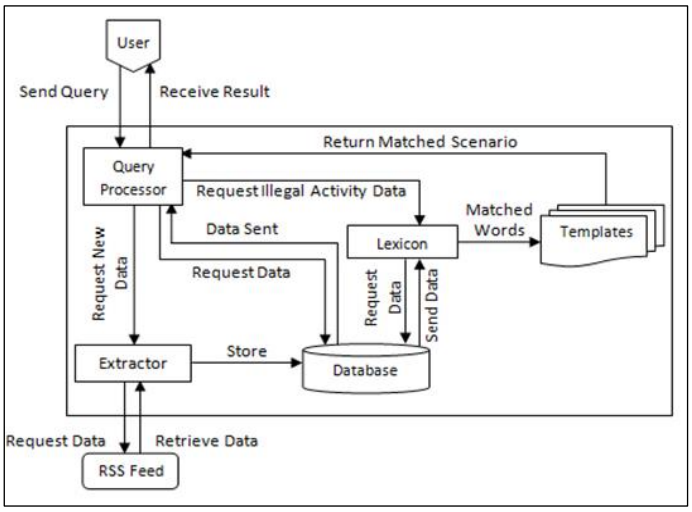

Fig. 2. Component Diagram for the Proposed FDBs-IDS

A. RSS FEEDS

The FDBs-IDS is capable of pulling and processing Really Simple Syndication (RSS) feeds in real time. RSS feeds are structured in Extensible Markup Language (XML) format. The FDBs-IDS extracts information through the tags in the RSS files and populate the FDBs-IDS database. This extraction and preprocessing is done through the extractor component which will be explained in the following section.

B. EXTRACTOR

The extractor component is the connection between the FDB's RSS feeds and the FDBs-IDS database. The extractor extracts comments, stock codes, times and dates and user ids from the FDBs RSS feeds then store them in the FDBs-IDS database.

C. FDBS-IDS DATABASE

The database behind the FDBs-IDS had been created for two purposes:

1. Storing user's comments, stock names, dates and times those have been extracted using the extractor.
2. Storing the lexicon (combination of known words and phrases). The words and phrases will be used in the IE templates explained in the following section.

D. LEXICON & IE TEMPLATES

The lexicon contains words and phrases associated with irregular activities on financial discussion boards. These words will be used either on their own or as part of more specific representations called IE Templates. The method for collecting keywords and phrases in order to populate the IE templates was through firstly through comprising lists of words and phrases used in previous published works [5, 17, 18]. Secondly, experts in FDB analysis provided a list of phrases/ words which were then added into the initial lists. Finally, the complete word and phrase list was reviewed by a further expert in financial fraud.

Therefore IE Templates are created based on concise representations of what constitute an irregular activity or potentially illegal activity such as pump and dump; simple example is shown in figure 3. The current IE templates deal with two irregular and potentially illegal activities which are pump and dump and announcing insider information. Words relating to each IE template are then stored into the lexicon ready to be used with the system. Then they can be used on their own or grouped according to template structure. The example shown in figure 4 shows highlighted words connected to words in the lexicon in which then triggered an IE template to be fired and thus highlight irregular and potentially illegal activity. The lexicon and IE templates are designed to allow them to be used for a wider range of activities as well in real time such as either mentioned in comments or within a larger template based matching.

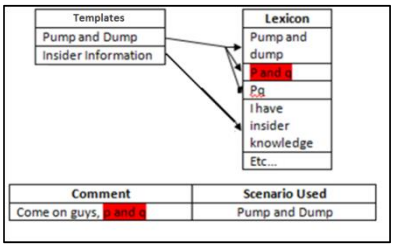

Fig. 3. Example Simple Template Structure

E. QUERY PROCESSOR

This component in the FDBs-IDS architecture synchronizes the user requests and the internal components within the FDBs-IDS. Figure 4 shows the functionality of the query processor. For example, it can be used to request real time data by requesting data from live RSS feeds. Also, it can be used to request and process data for offline analysis. In addition, it can be used to deal with the IE template filling using the known words and phrases associated with irregular and potentially illegal activity discussed earlier.

VI. FDBS-IDS IMPLEMENTATION

There have been a number of prototype systems created since our initial research in 2012 [10]. The FDBs-IDS system reported in this paper has been designed to be user friendly and more flexible based on feedback received on our previous prototype [10]. In addition, currently further research and development are being done. Interacting with the FDBs-IDS system is very simple, in which in few clicks, results on irregular and potentially illegal activity will be collected in real time, analyzed and returned. Users can select a template

to work with, and then click on analyse. Results then returned, highlighting any words associated with the selected template. Figure 5 below shows simple screenshot where the Pump and Dump template has been selected.

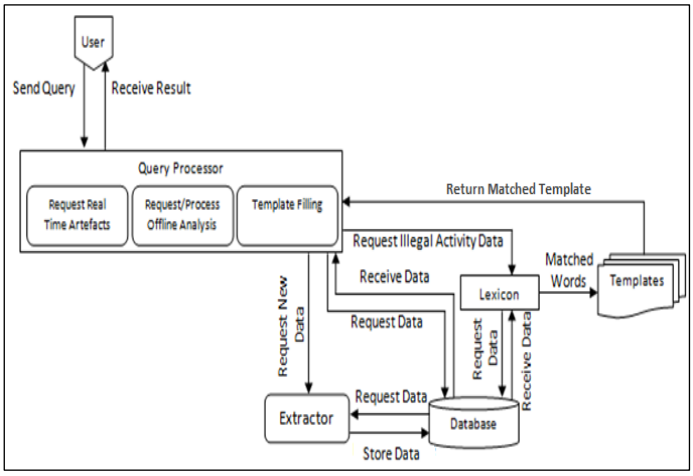

Fig. 4. Query Processor

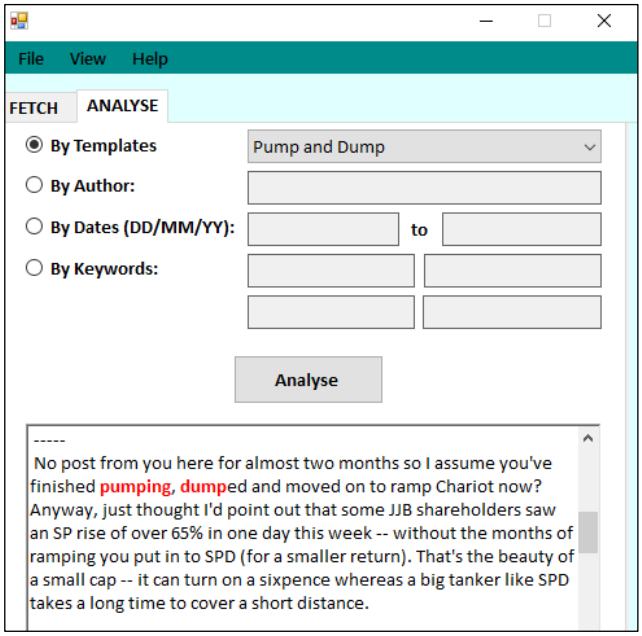

Fig. 5. Screenshot of FDBs-IDS System

In addition the FDBs-IDS will allow the user to perform the following three types of functions:

- Configure new templates and update current templates with new keywords and phrases.
- Highlight irregular and potentially illegal activity; this is through the use of IE templates on a specific user, a specific stock code, or on all FDBs data retrieved.
- Free search: allow the user search through the comments on a specific user, a specific stock code, or on all FDBs data retrieved as shown in figure 5.

VII. Experimental Methodology and Results

In order to validate the system, an experiment was conducted in which we have designed two templates which are the Pump and Dump and Insider Information. The overall statistics shows that on daily basis there were about 3000 - 5000 comments created only on one FDB [2]. The overall results shown in figure 6 show that from the comments collected on daily basis and analyzed against the two templates; there were a number of comments contain irregular or reveal a potentially illegal activity. The FDBs-IDS system flagged 4 potential pump and dump activities and 2 potential insider information activities through automatic detection. This is done a matter of minutes compared to a human operator in which this could take weeks.

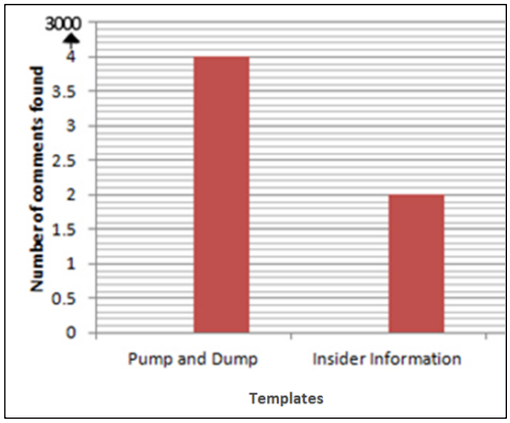

Fig. 6. Results for two templates

VIII. Conclusions

This paper has proposed a novel real time methodology for highlighting irregularities on FDBs to do this a novel FDBs-IDS was developed and tested. The FDBs-IDS prototype architecture uses IE as its main component to extract comments from FDBs to fill predefined templates for the analysis of potential irregularities or fraudulent activities. Results returned indicated that from 3000 comments made in one single day in one FDB alone, 6 activities were flagged as irregular or potentially discussing illegal activity. Over the course of one year, the FDBs-IDS could flag over 2,000 such comments from one single FDB. Therefore; this is significant number irregularities or potentially illegal activities and the FDBs-IDS will be a valuable tool to be used for criminal detection on the cyberspace. In addition, the FDBs-IDS was able to identify potential irregularities within minutes where human experts might spend weeks reading the FDBs blogs to identify irregularities. This research is currently looking at extending the extraction process to include more than one FDB and in addition to include more artifacts into account when reasoning about a potential irregularity or potential illegal activity on FDBs.

REFERENCES

[1] Interactive Investor, 2016. [Online]. Available: http://www.iii.co.uk. [Accessed 03 March 2016].

[2] London South East Limited, 2016. [Online]. Available: http://www.lse.co.uk. [Accessed 03 March 2016].

[3] Advfn PLC, "Advfn PLC," 2016. [Online]. Available: http://uk.advfn.com. [Accessed 03 March 2016].

[4] D. Leinweber and A. Madhavan, "Three Hundred Years of Stock Market Manipulations," 2001.

[5] J. Campbell and D. Cecez-Kecmanovic, "Communicative practices in an online financial forum during abnormal stock market behavior," *Information & Management*, vol. 48, no. 1, January 2011.

[6] J.-Y. Delort, B. Arunasalam, H. Leung and M. Milosavljevic, "The impact of manipulation in internet stock message boards," *International Journal of Banking and Finance*, vol. 8, no. 4, 2012.

[7] I. Alić, "Supporting Financial Market Surveillance: An IT Artifact Evaluation," in *BLED 2015 Proceedings*, 2015.

[8] P. Barnes, "Stock market scams, shell companies, penny shares, boiler rooms and cold calling: The UK experience," *International Journal of Law, Crime and Justice*, pp. 1-15, 2016.

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

[10] E. Knott and M. Owda, The detection of potentially illegal activity on financial discussion boards using information extraction, London: 2012 International Conference on Cybercrime, Security and Digital Forensics, 2012.

[11] L. Delamaire, H. Abdou and J. Pointon, "Credit Card Fraud and Detection Techniques: a Review," *Banks and Bank Systems*, vol. 4, no. 2, 2009.

[12] T. H. Roh, "Forecasting the volatility of stock price index," *Expert Systems with Applications*, vol. 33, no. 4, 2007.

[13] E. Appelt and D. Israel, "Introduction to Information Extraction," 1999.

[14] R. Mooney and R. Bunescu, "Mining knowledge from text using information extraction," pp. 3-10, 2005.

[15] R. Bunescu, R. Ge and E. Moone, "Comparative experiments on learning information extractors for proteins and their interactions.," 2005.

[16] FBI, 2012. [Online]. Available: https://www.fbi.gov/stats-services/publications/financial-crimes-report-2010-2011. [Accessed 21 March 2016].

[17] M. Lewis, "Jonathan Lebed's Extracurricular Activities," 25 February 2001. [Online]. Available: http://www.nytimes.com/2001/02/25/magazine/jonathan-lebed-s-extracurricular-activities.html?pagewanted=all&src=pm. [Accessed 24 March 2016].

[18] A. Riem, "Cybercrimes Of The 21st Century: Crimes against the individual," vol. 2001, no. 6, 2001.

[19] C. &. M. M. Chelba, "Information Extraction suing the structured language model," in *Proceedings of the 2001 Conference on Empiricla Methods in Natural Language Processing*, Pittsburgh, PA, USA, 2000.

[20] W. Antweiler and M. Frank, "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," *The Journal of Finance*, vol. 59, no. 3, 2004.

[21] S. Sabherwal, S. Sarkar and Y. Zhang, "Do Internet Stock Message Boards Influence Trading? Evidence from Heavily Discussed Stocks with No Fundamental News," *Journal of Business Finance & Accounting*, vol. 38, no. 9 & 10.

[22] H. Y. Limanto, N. N. Giang, V. T. Trung, N. Q. Huy, J. Zhang and Q. He, "An Information Extraction Engine for Web Discussion Forums," Singapore, 2005.

[23] M. Costantino, R. G. Morgan, R. J. Collingham and R. Garigliano, *Natural Language Processing and Information Extraction: Qualitative Analysis of Financial News Articles*, February 1997.

[24] M.-F. Moens, Information Extraction: Algorithms and Prospects in a Retrieval Context, Great Britain: Dordrecht, 2006, pp. 1-.

[25] J. Mena, in *Investigative Data Mining for Criminal and Security Detection*, Butterworth-Heinemann, 2003, pp. 3-4, 8, 126.

[26] R. Caruana and P. Hodor, in *High Precision Information Extraction, Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000.

[27] J. Srivastava, P. Desikan and V. Kumar. [Online]. Available: http://www.ieee.org.ar/downloads/Srivastava-tut-paper.pdf.