The detection of fraud activities on the stock market through forward analysis methodology of 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 on the FDBs. Popular public FDBs are prone to be used as a medium to spread false financial information due to having a larger group of audiences. Although online forums, in general, are usually integrated with anti-spam tools such as Akismet, moderation of posted contents heavily relies on human moderators. Unfortunately, popular FDBs attract many comments per day which realistically prevents human moderators from continuously monitoring and moderating possibly fraudulent contents. Such manual moderation can be extremely time-consuming. Moreover, due to the absence of useful tools, no relevant authorities are actively monitoring and handling potential financial crimes on FDBs. This paper presents a novel forward analysis methodology implemented in an Information Extraction (IE) prototype system named FDBs Miner (FDBM). This methodology aims to detect potentially illegal comments on FDBs while integrating share prices in the detection process as this helps to categorise the potentially illegal comments into different risk levels for investigation priority. The IE prototype system will first extract the public comments and per minute share prices from FDBs for the selected listed companies on London Stock Exchange (LSE). In the forward analysis process, the comments are flagged using a predefined Pump and Dump financial crime related keyword template. By only flagging the comments against the keyword template yields an average of 9.82% potentially illegal comments. It is unrealistic and unaffordable for human moderators to read these comments on a daily basis in long run. Hence, by integrating the share prices' hikes and falls to categorise the flagged comments based on risk levels, it saves time and allows relevant authorities to prioritise and investigate into the higher risk flagged comments as it can potentially indicate real Pump and Dump crimes on FDBs.**

Keywords— Financial Discussion Boards; Fraud Detection; Crime Prevention; Financial Crimes; Pump and Dump; Text Mining; Information Extraction

I. INTRODUCTION

Given the freedom of speech on the Internet, there are many online forums where like-minded people can hold conversations in the form of posted messages. Financial Discussion Boards (FDBs) allow investors to exchange knowledge, information, experience and opinions about the investment opportunities. There is a various popular share price based FDBs in the UK which specifically allows investors to discuss share prices. These FDBs include the London South East [1], Interactive Investor [2] and ADVFN [3].

Although online forums seem to be a useful source of information, not all information shared on the forums is accurate or truthful. Anti-spam plugins such as Akismet [4] are usually the default tools integrated on most online forums to filter and prevent spammers from registering or posting spam messages. However, such tool does not moderate the meaning of a content. Similarly, a forum moderator handles only the offensive and/or prohibited contents such as racism, sexism, hatred, foul languages, third party advertisements and so on. There are very little to no measures taken by forum moderators and external authorities to monitor and detect potential crimes on the FDBs, such as comments indicative of Pump and Dump (P&D). Such manual moderation on FDBs requires an enormous amount of time and effort, which is not feasible in long run.

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. Research from recent years highlighted that the comments on FDBs were found manipulative and positively related to the market returns, volatility and trading volumes [5, 6, 7, 8, 9]. However, there is very little attempt [10, 11] made to build tools

for monitoring and detection of potential financial crimes on share price based FDBs.

FDBs contain semantically understandable artefacts (i.e. FDBs' artefacts that can be processed by computers) such as stock ticker names, date, time, prices, comment author usernames and comments. Information Extraction (IE) techniques are used in this research to extract these artefact data. IE is defined as the process of extracting information automatically into a structured data format from an unstructured or semi-structured data sources [12]. IE has been used in other areas such as accounting [13] and search engine [14]. However, other than the initial work described in [11] and [15], there is very little usage of IE techniques in FDBs' financial crimes related research.

During the detection of potentially illegal comments in [15], share prices were not considered. Hence, the novel methodology introduced in this paper will be used to flag all the potentially illegal comments while integrating the share prices into the detection process. This methodology is implemented in an IE prototype system named FDBs Miner (FDBM). FDBM will start by analysing all the comments against a predefined P&D IE keyword template. Then, 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 of 5%, 10% and 15% compared to the "base price". Finally, it appends the price hike threshold labels to these flagged comments. The main contribution of this paper is to introduce a novel methodology that will flag potentially illegal comments as well as categorise these comments based on the level of risks. This can greatly benefit the relevant authorities to prioritise and investigate into the potentially illegal comments according to risk levels.

Section II reviews the examples of past financial crimes on share price based FDBs. Section III introduces Information Extraction (IE) and its usage in FDBM. Section IV presents an architecture overview of FDBM. This followed by Section V which describes the novel forward analysis methodology and the experimental results in Section VI. Lastly, Section VII concludes the research findings.

II. PUMP AND DUMP (P&D) CRIMES ON FDBS

P&D crimes are normally committed through various mediums such as discussion boards, word of mouth, social media, emails and so on. The following are a few examples of the popular share price based P&D financial crimes:

- 15-year-old Jonathan Lebed was the first minor involved in a stock market fraud in 2000 [16]. 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) [16, 17, 18].

- 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 [17].

- In 2009, eight participants were charged by Security Exchange Commission (SEC) for being involved in penny stock manipulation [19]. These criminals met each other through InvestorsHub (now owned by ADVFN [3]), a popular penny stock message board, and carried out the P&D scheme throughout the year of 2006 and 2007.

Based on the above FDBs related P&D crimes, there is a clear and persistent need to create methods and tools to detect potentially illegal contents on share price based FDBs in real time.

III. INFORMATION EXTRACTION (IE)

IE is the process of extracting information such as text from unstructured or semi-structured data sources into a structured data format [12]. Soderland [20] suggested that there is a need for systems that extract information automatically from text data. IE systems are knowledge-intensive [20] 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 end users of a system [21].

IE is used in this research to automatically extract information from an unstructured or semi-structured data source (such as FDB comments and share prices) into a structured data format (i.e. FDBs dataset). The IE prototype system in this research can display a summary of information from several interlinked sources (i.e. FDB comments and share prices) allowing filtering of potentially illegal comments to take place.

IV. AN ARCHITECTURE OVERVIEW OF FDBS MINER (FDBM)

This section presents the FDBM architecture which consists of five key components. These key components are the data crawler, data transformer, FDB dataset (FDB-DS), IE keyword template and the forward analyser. In general, FDBM will first collect data, then transform unstructured and semi-structured data into fully structured data which kept in the FDB-DS. The IE keyword template is for the use with the forward analyser. The novel methodology introduced in this paper is made functional in the forward analyser component.

Figure 1 provides an architecture overview of the FDBM prototype system. Each component in the architecture diagram is described in the following sections.
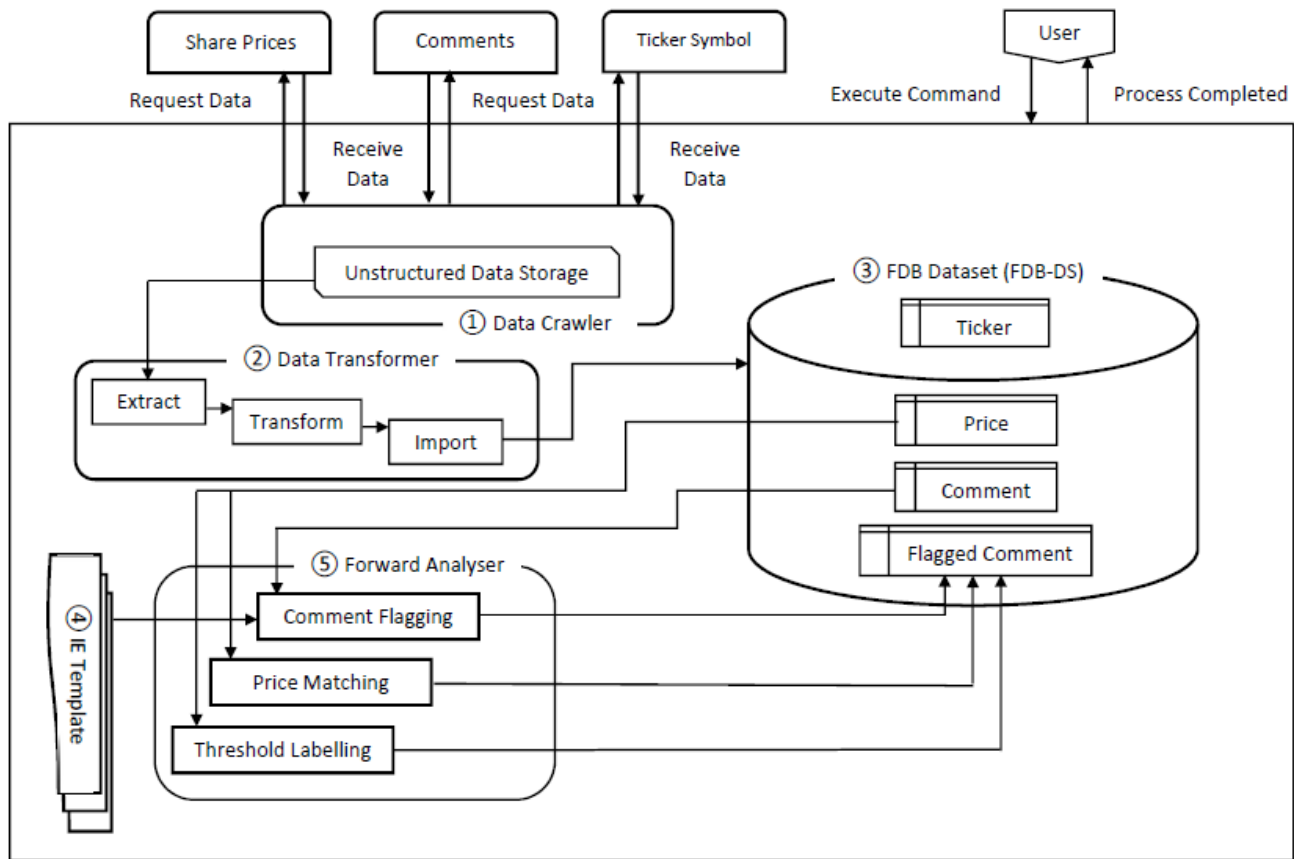
Fig. 1. Architecture Overview Diagram.

A. Data Crawler

The data crawler is responsible for automatically collecting unstructured data from the three FDBs (i.e. LSE [1], III [2] and ADVFN [3]) at different time intervals for 12 weeks (23rd September 2014 to 22nd December 2014). A total of 941 ticker symbols (i.e. unique abbreviations of companies listed on the stock market), 507,970 FDB comments and 28,980,465 price figures were collected.

B. 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.

C. FDB Dataset (FDB-DS)

After the collected data is being processed by 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. The FDB-DS is also responsible to store additional data produced from research analysis.

D. IE Keyword Template

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 be easily modified whenever needed. The IE keyword template consists of a series of keywords and phrases

that were thoroughly researched [22, 23, 24, 25] and has been validated by experts in the relevant field. The IE keyword template will be used by the forward analyser for the comments flagging process.

E. 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 will be further discussed in Section V.

V. FORWARD ANALYSIS METHODOLOGY

This section introduces the novel forward analysis methodology. This methodology flags and filters the potentially illegal P&D comments using P&D keyword template. It also integrates the share prices in the analysis process in order to categorise the flagged comments into different risk levels. This allows relevant authorities to investigate into the flagged comments more realistically in terms of time and efforts.

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

A. Comments Flagging

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

B. Prices and Comments Flagging

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.

C. Comments Threshold Labelling

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, the forward analyser first calculates all the ± 2 days' per minute prices against the "base price" of each flagged comment.

The threshold labelling rules are listed 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".

VI. FORWARD ANALYSIS 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 1). These flagged comments took up 9.82% of the total comments.

TABLE I. 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 2 and visualised in Figure 2.

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

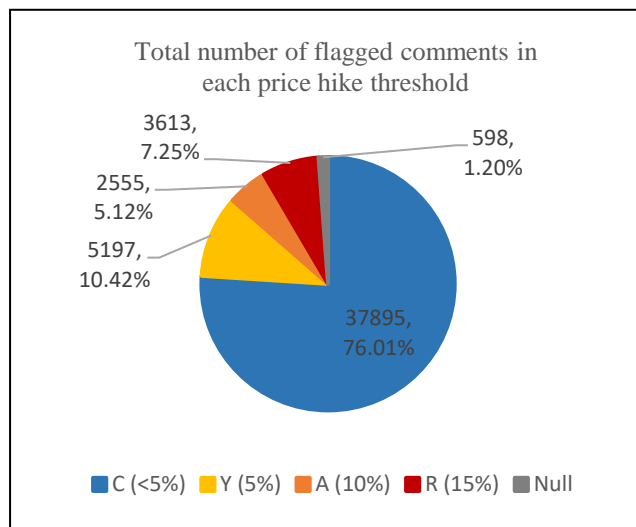| Threshold	Total	Percentage
C (<5%)	37,895	76.01%
Y (5%)	5,197	10.42%
A (10%)	2,555	5.12%
R (15%)	3,613	7.25%
Null	598	1.2%
Grand Total	49,858	100%

Fig. 2. Total number and percentage of each threshold.

The time taken in this analysis process is long, however, this is only due to the significant amount of data being processed and analysed. If the prototype system and methodology are used in real time in real world scenario, it can significantly reduce the time, effort and cost of monitoring and detecting P&D crimes on FDBs.

VII. CONCLUSION

This paper has introduced a novel methodology 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 analysis methodology integrates 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 methodology 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.

VIII. REFERENCES

[1] London South East Limited, "London South East" [Online]. Available: http://www.lse.co.uk, September 2017

[2] Interactive Investor Plc., "Interactive Investor" [Online]. Available: http://www.iii.co.uk, September 2017

[3] ADVFN PLC, "ADVFN" [Online]. Available: http://uk.advfn.com, September 2017

[4] Akismet, "Akismet" [Online]. Available: https://akismet.com, September 2017

[5] 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.

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

[7] 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.

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

[9] 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.

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

[11] 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.

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

[13] 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.

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

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

[16] Lewis, M., "Jonathan Lebed's Extracurricular Activities," The New York Times [Online]. Available: http://www.nytimes.com/2001/02/25/magazine/jonathan-lebed-s-extracurricular-activities.html?pagewanted=all&src=pm, September 2017.

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

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

[19] US Security and Exchange Commission, "SEC Charges Eight Participants in Penny Stock Manipulation Ring" [Online]. Available: http://www.sec.gov/litigation/litreleases/2009/lr21053.htm, September 2017.

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

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

[22] 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.

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

[24] 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.

[25] 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.