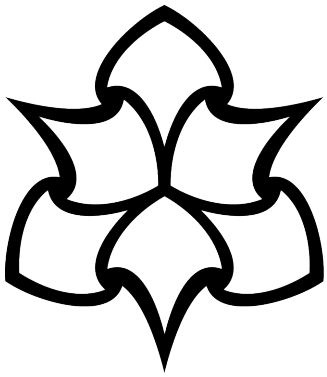

A Financial Crime Analysis Methodology for Financial Discussion Boards using Information Extraction Techniques

PEI SHYUAN LEE

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

School of Computing, Mathematics and Digital Technology
Manchester Metropolitan University

April 2018

Declaration

I declare that no portion of the work referred to in the thesis has been previously submitted for a degree or qualification at any other university or other institute of learning.

PeiShyuanLee

Pei Shyuan Lee

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 be used as a medium for spreading misleading financial information due to having larger audience groups. Moderation of posted content heavily relies on manual tasks. Unfortunately, the daily comments volume received on popular FDBs realistically prevents human moderators or relevant authorities from proactively monitoring and moderating possibly fraudulent FDB content as it is extremely time-consuming and expensive to manually read all the content.

This thesis presents a financial crime analysis methodology (which is comprised of novel forward analysis and novel backward analysis methodologies) implemented in a template-based Information Extraction (IE) prototype system, namely FDBs Miner (FDBM). The methodologies aim to detect potentially illegal Pump and Dump (P&D) activities on FDBs with the integration of per minute share prices in the detection process. This integration can reduce false positives during the detection as it categorises the potentially illegal comments into different risk levels for investigation purposes.

P&D is a well-known financial crime that happens through different methods including FDBs. P&D happens when fraudsters deceive investors into buying stocks by spreading misleading information. FDBM extracts a company's ticker symbol (i.e. a unique symbol that represents and identifies each listed company on the stock market), comments and share prices from FDBs based in the UK for experimental purposes. Results from both forward and backward analysis experiments show that the two novel methodologies can aid relevant authorities in the detection of potentially illegal activities on FDBs. Semantic Textual Similarity (STS) experiments have also shown that the approach could be adopted in the process of detecting potentially illegal activities on FDBs.

Acknowledgement

I am indebted and thankful to my parents who supported me financially to pursue my PhD research in the UK. I am also wholeheartedly grateful to my supervisors, Dr Majdi Owda and Dr Keeley Crockett, who did not give up on me when I did not progress well due to personal matters. I appreciate their efforts in always pulling me back on track, following up my work and guiding me throughout the entire process of my PhD research journey, right up to the last minute. Also, thank you to the anonymous expert in the field who validates the financial crime keyword template and was part of one of the conducted experiments. I would also like to give special thanks to my best friend for giving me a lot of positive vibes, listening to my problems and being patient with me for years. Thank you to my best housemate for two years, who took care of my meals when I was stressed out due to my PhD research. Lastly, I also thank my other Malaysian friends who were there for me in the first two years of my PhD journey.

Table of Contents

List of Abbreviations

AIM	Alternative Investment Market
CSV	Comma-Separated Values
EMA	Exponential Moving Average
FCA	Financial Conduct Authority
FDB	Financial Discussion Boards
FDB-DS	Financial Discussion Boards Dataset
FDBM	Financial Discussion Boards Miner
FSA	Financial Services Authority
FTSE	Financial Times Stock Exchange
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IE	Information Extraction
II	Insider Information
III	Interactive Investors FDB
IR	Information Retrieval
KE	Knowledge Engineer
LSA	Latent Semantic Analysis
LSE	London Stock Exchange
LSE-FDB	London South East Financial Discussion Board
MA	Moving Average
MAD	Market Abuse Directive

MAR	Market Abuse Regulation
NLTK	Natural Language Toolkit
P&D	Pump and Dump
PRA	Prudential Regulatory Authority
RNS	Regulatory News
RSS	Really Simple Syndication
SEC	Security Exchange and Commission
SMA	Simple Moving Average
STS	Semantic Textual Similarity
SVD	Singular Value Decomposition
URI	Uniform Resource Identifier
URL	Universal Resource Locator
WMA	Weighted Moving Average
XML	Extensible Markup Language

Glossary

Broker Ratings – Suggestions or recommendations by stock brokers whether a stock is worth buying or selling.

Director Deals – Stocks bought and sold by the directors of a listed company.

Financial Diary – Financial details such as company's turnover, profit, market capital and so on.

Penny Stock – Stocks that are usually valued at less than a dollar.

Semantically Understandable Artefacts – FDBs' artefacts that can be processed by computers

Ticker Symbol – A unique symbol that represents and identifies each listed company on the stock market.

List of Publications

The following papers have reported some of work related to this thesis:

1. Lee, P. S., Owda, M. and Crockett, K. (2014) 'A Financial Crime Analysis Methodology for Financial Discussion Boards using Information Extraction Techniques.' *7th Manchester Metropolitan University Postgraduate Research Conference*, Manchester Metropolitan University, Manchester.
2. Owda, M., Lee, P. S. and Crockett, K. (2017) 'Financial Discussion Boards Irregularities Detection System (FDBs-IDS) using Information Extraction.' *Intelligent Systems Conference 2017*. London, 7th-8th September 2017.
3. Lee, P. S., Owda, M., & Crockett, K. (2018) 'The detection of fraud activities on the stock market through forward analysis methodology of financial discussion boards.' *Future of Information and Communications Conference*. Singapore, 5th-6th April 2018.
4. Lee, P. S., Owda, M. and Crockett, K. (2018) 'Methodologies for resolving false positives during the detection of fraudulent activities on the stock market through forward and backward analysis of financial discussion boards.' *International Journal of Advanced Computer Science and Applications*, 9(1).

List of Figures

List of Tables

Chapter 1. Introduction

This thesis describes the development of a Stock Market Surveillance Prototype System, namely Financial Discussion Boards Miner (FDBM), which implements a set of novel methodologies in an attempt to detect potential financial crimes on share price based Financial Discussion Boards (FDBs) in the UK. FDBs contain numerous semantically understandable artefacts (i.e. FDBs' artefacts that can be processed by computers) which will be used for this research.

Given the freedom of speech on the Internet, it has become the number one source of information for unlimited things. Unsurprisingly, this includes financial discussions by investors and traders. There are many online forums where likeminded people can hold conversations in the form of posted comments (vBulletin Solutions Inc., 2017). FDBs allow investors to exchange knowledge, information, experience and opinions about investment opportunities.

Such FDBs are not moderated by relevant authorities nor strictly moderated by forum owners and moderators. Due to the lack of manpower or systems to moderate posted content automatically, share price based FDBs are open to abuse by fraudsters. Financial crime like Pump and Dump (P&D) is the predominant crime that can happen on FDBs. Typically, fraudsters post false news, usually in an encouraging or positive manner about particular stock on the FDBs after buying in a huge amount of stock at a very low price. This attracts genuine investors into buying the stock. Once the stock price is pumped up due to high demand, the fraudsters sell off a huge portion of stock. The stock price is immediately dumped, leaving the deceived investors with losses. Penny stocks (i.e. stocks valued at less than one dollar) are often used in P&D crimes as it is a lot easier for fraudsters to pump the price and gain a significant amount of profit almost effortlessly.

The stock exchange used for this research is the London Stock Exchange (LSE, 2017) and the share price based FDBs in the UK are ADVFN (ADVFN, 2017), London South East (LSE-FDB, 2017) and Interactive Investors (III, 2017). Based on thorough research

and observations by the thesis author, these three FDBs appear to be the most popular and active FDBs based in the UK. The authors of previous work (Delort et al., 2011; Knott and Owda, 2012; Leung and Ton, 2015) have also picked the most popular FDBs based in the countries in which their research is conducted as popular and active FDBs attract more members, thus there is a higher chance of financial crimes. It appears that very little Information Extraction (IE) research has been conducted in relation to the analysis of potentially illegal activities on share price based FDBs. As a result, this research explores the potential usage of IE techniques, Moving Average (MA) techniques and Semantic Textual Similarity (STS) in the FDBM prototype system. The solution presented in this research could significantly influence the way share price based FDBs are regulated in the future.

This chapter describes the motivation for this study, a list of the aim and objectives, a summary of contributions and the thesis outline.

1.1 Background and Motivation

The development of the Financial Discussion Boards Miner (FDBM) prototype system presented in this research adopted techniques in three research areas, namely Information Extraction (IE), Moving Average (MA) and Semantic Textual Similarity (STS).

The increased freedom of speech on the internet means an increased chance of delivering misleading information regardless of whether it is with the intention or not. Discussion boards are popular since they represent a place for topical discussions by likeminded people. Being part of the discussion boards, share price based FDBs allow investors to discuss investment opportunities, exchange and share knowledge. They have also become a place for fraudsters to commit financial crime like P&D by deceiving others into buying stock through the spreading of misleading information.

One of the problems with such FDBs is that the investors do not know the exact intention of another poster (Ackert et al., 2016). Fraudsters can easily pretend they are experienced investors and deceive others into buying stock. P&D fraudsters on

FDBs usually buy a huge amount of stock before spreading misleading information on FDBs, usually accompanied by positive sentiment such as “buy now”. When the price of the stock has successfully been pumped up, the fraudsters “dump” their stock and get away with huge profits. This leaves the victims with losses. Another problem with most of the FDBs is that, since FDBs are used for financial discussions, forum moderators do not moderate the financial related sentiments or opinions of investors (Ackert et al., 2016). Even if the FDB comments are meant to be moderated, it is unrealistic for the forum moderators to read through all the comments on a daily basis, especially on popular share price based FDBs.

Financial crime related research in the past has tended to focus on prediction through trading volume and quantified content from the Internet (Jin et al., 2016). Wysocki (1999) was one of the first authors to find a positive relationship between the trading volume and the number of posted messages on Yahoo! message boards. Other established research in the field such as Antweiler and Frank (2004), Cook and Lu (2009) and Bettman et al. (2011) also focuses on finding a relationship between trading volume and posted content volume from FDBs.

Delort et al. (2011) attempted to automate the moderation of Online Discussion Sites (ODSs) using their novel classification technique with the incorporation of a partially labelled corpus, i.e. comments that were moderated and labelled by forum moderators on an Australian FDB, namely HotCopper (HotCopper, 2017). Their classification was able to moderate the comments into various categories, not limited to ramping (i.e. P&D), such as flaming, profanity, spamming and so on. However, according to the authors, the misclassification remains too significant. Furthermore, the share prices were not taken into account during the moderation of ODS content.

Other than the initial work proposed by Knott and Owda (2012), there has been no other research attempt using IE techniques in relation to the monitoring and detection of potential P&D activities on share price based FDBs in the UK. The initial work proposed a template-based IE prototype system that can flag potentially illegal FDB comments. However, the initial work was a fairly basic system as it did not take share prices into account.

To the knowledge of the thesis author, no other detection or moderation tool has been built for detecting potentially illegal P&D activities on share price based FDBs in the UK by taking share prices into account during the detection process.

The motivation for this research came from the need for a stock market surveillance prototype system that can detect the potentially illegal P&D activities on share price based FDBs. P&D is nothing new; it is still happening today, through many methods which include the use of share price based FDBs as they allow investors to communicate. As long as the stock exchanges are alive, so is P&D. P&D happens especially around the penny stocks (Barnes, 2017) because these stocks are usually listed under the index, such as FTSE (pronounced as "Foot-sie") AIM All-Share in the London Stock Exchange (LSE, 2017) that has more flexible regulation than the main indices such as the FTSE-100 and FTSE-250. The reason for having more flexible regulation for such an index is because it allows new and smaller companies to be part of the stock market. Unfortunately, this opens the door to fraudsters.

This project attempts to address the following research question:

> Can IE techniques, MA techniques and STS be used to develop novel methodologies to automatically detect potential financial crimes on share price based FDBs?

The testable hypothesis is listed as follow:

> H_0: Potential Pump and Dump financial crime activities on financial discussion boards (FDBs) cannot be detected through the use of IE techniques, MA techniques and STS.
>
> H_1: Potential Pump and Dump financial crime activities on financial discussion boards (FDBs) can be detected through the use of IE techniques, MA techniques and STS.

1.2 Research Aim and Objectives

The aim of this research is to investigate the suitability of Information Extraction (IE) techniques, Moving Average (MA) techniques and Semantic Textual Similarity (STS); and, the combined use of these techniques in developing a template-based Information Extraction (IE) stock market surveillance prototype system that is capable of automatically detecting potential financial crimes such as Pump and Dump (P&D) on share prices based Financial Discussion Boards (FDBs).

In order to address the research question, this research can be divided into a number of objectives. The objectives of this research are listed as follow:

1. To investigate the suitability of IE techniques, MA techniques and STS usage in FDB related research and to establish the background work in the field. This is to determine the state of the art in the fields.
2. To identify the semantically understandable artefacts on FDBs such as stock ticker names, date, time, usernames, comments and more. These artefacts will be employed for the experiments conducted in this research using the prototype system.
3. To research and evaluate financial crime related keywords used in FDBs and elect appropriate keywords for the use of novel IE keyword template creation in conjunction with the semantically understandable artefacts.
4. To capture 12 weeks of message board comments traffic and share prices from FDBs in order to construct an FDB dataset (FDB-DS) for experiment analysis and evaluation.
5. To develop a novel methodology for a template-based IE prototype system by using IE techniques.
6. To develop an automated prototype system based on the designed methodology in an attempt to perform real-time financial surveillance and reduce time consumption when moderating FDB comments.
7. To apply statistical techniques such as Simple Moving Average (SMA), Weighted Moving Average (WMA) and Exponential Moving Average (EMA) on the prices in order to highlight rises and falls in price regardless of the availability of flagged comments.

8. To integrate the use of STS in the prototype system so that the textual comments can be compared to the novel IE keyword template.
9. To evaluate the prototype system by conducting experiments using the constructed FDB-DS.

1.3 List of Contributions

The principal research contributions are listed as follow:

1. A semi-automated stock market surveillance prototype system, namely FDBs Miner (FDBM), developed based on the novel methodologies devised in Contribution 5, that is capable of analysing FDB comments and share prices simultaneously in order to flag potentially illegal activities on the share price based FDBs.
2. A crawler component in the FDBM prototype system that is capable of crawling semi-structured data from share price based FDBs automatically.
3. A data transformer component in FDBM that can pre-process and transform the FDB related semi-structured data, collected in Contribution 2, into structured data.
4. A novel FDB dataset (FDB-DS) that contains FDB artefacts data such as ticker symbols, FDB comments, share prices, broker ratings and director deals, which belong to all of the 941 companies (picked from FTSE-100 and FTSE AIM All-Share indices) on the London Stock Exchange (LSE, 2017). These data can be used for further research work and extension of the prototype system.
5. Two novel methodologies (i.e. forward analysis and backward analysis) that are formulated based on the IE and MA techniques for the detection of potentially illegal activities on FDBs.
    - Forward analysis can perform the flagging comments against the list of keywords, phrases and sentences in the P&D IE keyword template. It is also capable of calculating the ±2 days of share prices against the "base price" of a flagged comment and appending price hike thresholds to these flagged comments.

- The backward analysis can calculate the moving averages of all the per minute share prices and highlight the abnormalities in price movements. It is also capable of labelling these price abnormalities backward to the flagged comments to further classify the flagged comments for investigation prioritisation and resolving false positives.

6. A predefined IE keyword template that was constructed based on P&D financial crime. The template contains keywords, phrases and sentences, that are commonly used by fraudsters on FDBs. The keyword template can be employed and expanded in a real-world scenario by the relevant authorities.
7. A prototype system component that can accept input of new financial crime IE keyword templates defined by relevant authorities; thus, FDBM is not only for detecting P&D financial crimes but also for detection of other potential financial crimes on FDBs.
8. The STS approach that is incorporated into the forward analysis can reduce false positives during the comment flagging process. Relevant authorities can investigate potentially illegal FDB comments based on different semantic similarity thresholds while performing both forward and backward analysis as proposed in the two novel methodologies described in Contribution 5.

1.4 Thesis Outline

This chapter presents an overview of the research background and motivation, research aim, research objectives and research contributions. Figure 1.1 illustrates a flowchart of the interrelation between the thesis chapters.

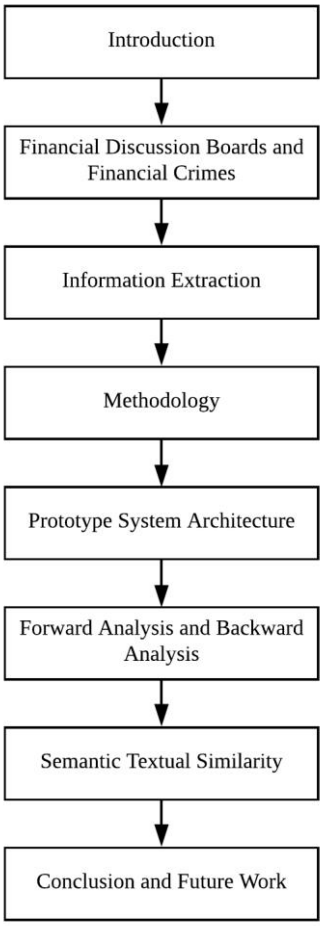

Figure 1.1 Research Chapter Flowchart

The thesis is organised into seven chapters. Chapter 2 provides an idea of the stock exchange in the UK, namely the London Stock Exchange (LSE, 2017). The same chapter also reviews all three share price based FDBs in the UK on which the stock market price data are collated from the LSE. While reviewing the share price based FDBs, the chapter also identifies the semantically understandable artefacts on FDBs. The stock market regulatory agencies in both the UK and the US are also introduced. The chapter continues with a discussion of existing popular financial crime cases that

have happened on FDBs which were mainly dealt with by the regulatory agency in the US. This is followed by a description of the existing research related to P&D financial crime which has happened through methods such as emails, social media and FDBs.

Next, Chapter 3 introduces and describes the three types of data structures that can be collected from the Internet. Following this, there is an introduction of the two fundamental classes of IE, namely the knowledge engineering approach and automatic learning approach. The knowledge engineering approach is also known as the rule-based approach as it relies on rules to be predefined by the experts in the domains in order for the systems to work out the outcomes. This approach is discussed as it is the approach being utilised by the template-based IE prototype system introduced in this thesis. An overview of Semantic Similarity Measures is also presented in the same chapter.

Chapter 4 introduces methodologies for the detection of potential financial crime activities on share price based FDBs. The methodology for identifying the FDBs and the semantically understandable artefacts is described. This is followed by a description of the overall methodology (Section 4.3) for the development and implementation of the prototype system, which is divided into seven phases (i.e. the implementation of a data crawler, data transformer, dataset for storing FDB data, construction of P&D IE keyword template, novel forward analyser, novel backward analyser and Semantic Textual Similarity).

A prototype system architecture overview is presented in Chapter 5. Section 5.3 to Section 5.8 in the chapter describe each of the components in the architecture that implements the 7-phase methodology. Lastly, the Graphical User Interface (GUI) for the data crawler, novel forward analyser and novel backward analyser is presented.

Chapter 6 describes a series of experiments for the novel forward analysis and backward analysis introduced in this research – two experiments for the forward analysis and another two for the backward analysis. The outcomes of the experiments are discussed. Chapter 7 also conducts experiments using the STS approach to test whether the STS approach can be adopted in the detection of

potentially illegal FDB activities. The outcomes of the experiments are also discussed. Finally, Chapter 8 concludes the thesis, highlights the contributions and discusses the potential future directions of this research.

Chapter 2. Financial Discussion Boards and Financial Crimes

2.1 Introduction

Over the years, the Internet has become a place for people to seek and share information and almost everything else that could possibly be presented in a digital format. The Internet does seem like a major source of information in many ways. However, not all information being shared online is accurate. Furthermore, not everyone undertakes the necessary due diligence prior to acting upon the information found on the Internet. With the freedom of speech on the Internet, many things can go wrong. One of these things is the spreading of fake financial information through Financial Discussion Boards (FDBs) on the Internet.

Unlike the traditional methods, false information does not just spread through word of mouth and spam emails anymore; it also spreads through FDBs since these are the places that gather the most investors. Share price based FDBs represent a place for investors to exchange stock related financial knowledge and discuss stock investment or trading related topics in the form of posted comments. These comments often involve investors' financial sentiments. However, some fraudsters may disguise themselves as investors and spread fake financial information or actively promote specific stocks, usually penny stocks (i.e. stocks that are valued at less than one dollar).

Each country has its own stock exchange that allows investors to trade and invest. Share price based FDBs are created based on these stock exchanges. There have been a fair number of studies conducted on the New York Stock Exchange (NYSE) (NYSE, 2017), NASDAQ (NASDAQ, 2017) and the Australian Stock Exchange (ASX) (ASX, 2017). However, there is little to no FDB related research which has been conducted on the London Stock Exchange (LSE) (LSE, 2017) for the infamous Pump and Dump (P&D) financial crime. This research examines the share price based FDBs that collate stock market data from the LSE. P&D happens when fraudsters intentionally and actively post comments like "This is the right time let's start pumping this share" after buying specific stocks at a low price. Novice investors often fall into such schemes. Once the

price hikes to a certain level, the fraudsters “dump” their stocks, earning them huge illegal profits while leaving others at a financial loss.

This chapter provides a comprehensive review of FDBs and how they have been used for financial crimes. The key problems and challenges of detecting such crimes on FDBs are also discussed.

2.2 London Stock Exchange

The LSE (LSE, 2017) is located in the city of London, UK. As of 11th April 2017, according to an infographic (Visual Capitalist, 2017), the LSE is one of the oldest stock exchanges (215 years old) and it is the third largest stock exchange in the world after the NYSE and NASDAQ. The LSE has a total market capital size of 6.187 trillion US dollars (Visual Capitalist, 2017) and has 3,041 listed public companies. Each listed company has a unique ticker symbol, i.e. a unique abbreviated name that represents and identifies each listed company on the stock market.

The LSE has seven main indices and each index is described below:

- **FTSE-100**

  This index consists of the first hundred companies with the highest market capitalisation listed on the LSE.
- **FTSE-250**

  This index consists of the 101st to the 350th largest companies listed on the LSE.
- **FTSE-350**

  This index consists of a combination of both the FTSE-100 and FTSE-250 listed on the LSE.
- **FTSE All-Share**

  This index consists of the combined indices of the FTSE-100, FTSE-250 and FTSE SmallCap (i.e. an index that consists of the 351st to 619th listed companies on the LSE).

- **FTSE AIM UK 50**

  This index consists of the 50 largest UK companies listed by the Alternative Investment Market (AIM). Note: AIM allows smaller companies to offer shares to the public with a more flexible regulatory system as compared to the main market like FTSE-100.

- **FTSE AIM 100**

  This index consists of the first hundred companies listed on the AIM. Other than non-UK companies, this index may also consist of UK companies listed on the FTSE AIM UK 50.

- **FTSE AIM All-Share**

  This index consists of all the UK and non-UK companies listed on the AIM.

In this research, the FDB comments on the FTSE-100 and FTSE AIM All-Share listed companies, from 23rd September 2014 to 22nd December 2014, have been chosen for the experiments. This allows for a comparison of whether potentially illegal P&D activities tend to happen on both indices, the FTSE-100 or FTSE AIM All-Share. As mentioned, companies that are listed on the FTSE AIM indices are smaller companies and have a more flexible regulatory system as compared to the main indices like the FTSE-100. Smaller (penny stock) companies tend to be abused by fraudsters for P&D crimes (Barnes, 2017).

2.3 Share Price Based Financial Discussion Boards (FDBs)

According to Lu et al. (2010), financial related content falls into three categories. The first category consists of forums, blogs or wikis. The second category contains news or research reports. The third category describes those firms which have a means by which investors can communicate. As Feldman et al. (2008) claimed, a discussion board is a place which can provide essential information such as consumers' sentiments towards a particular product.

This research extracts information from three share price based FDBs based in the UK. These FDBs are ADVFN (ADVFN, 2017), London South East (LSE-FDB, 2017) and

Interactive Investor (III, 2017) respectively. Each of these FDBs will contribute to the identification of semantically understandable artefacts and help shape the FDB dataset (FDB-DS) for the use of experiments in subsequent chapters.

2.3.1 ADVFN

ADVFN (ADVFN, 2017) is one of the largest share price based FDBs globally and it was founded in 1999. ADVFN does not just provide a discussion board for interactions between investors; it also provides several types of financial information relevant to the stock market. ADVFN is divided into many sections based on stock market regions across the globe such as the United States of America (USA), Canada, Brazil, India, Japan, Russia, Saudi Arabia, the United Kingdom (UK), Italy and so on. The UK version of ADVFN is used since the LSE stock market is being studied in this research. Figure 2.1 demonstrates an example of how various information on a specific share, for example, Lloyds (ticker symbol: LLOY), would appear on the ADVFN website.

Figure 2.1 Lloyds (LLOY) on ADVFN

In order to participate in the discussions and access other functions of the website like a personal portfolio, investors can sign up for free accounts on ADVFN. To access the historical per minute share prices, the user needs to pay for a monthly

subscription. ADVFN appears to be the only FDB that provides per minute share price history.

2.3.2 London South East

London South East (LSE-FDB, 2017) was established in 1997. It focuses only on the London Stock Exchange. Like ADVFN, LSE-FDB provides not only discussion boards for interaction but also other types of information related to the stock market. Figure 2.2 illustrates the webpage for how a ticker symbol such as LLOY would appear on the LSE-FDB website. In addition to most of the artefacts on ADVFN, LSE-FDB also provides broker ratings and director deal artefacts which ADVFN does not.

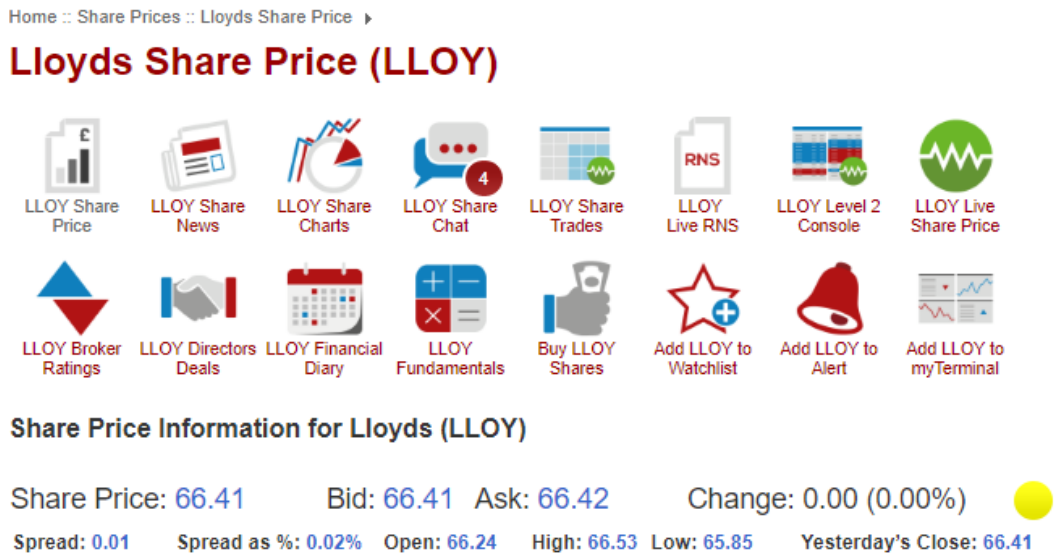

Figure 2.2 Lloyds (LLOY) on LSE-FDB

Similar to ADVFN, in order to access the live prices and certain premium features, LSE-FDB requires a member to subscribe to its monthly paid subscription. The only difference between ADVFN and LSE-FDB in terms of access to share prices is that ADVFN provides both live prices and downloadable historical prices whereas LSE-FDB provides only the live prices. It is more time consuming to capture the live prices in this research because there is a need to design a separate crawler component just to

collect the per minute share price. This is also risky because the process of data crawling might get interrupted due to unforeseen circumstances. Furthermore, although the owner of the LSE-FDB has expressed his interest in this research in an email reply (see Appendix E), he does not feel comfortable with this research crawling the live prices on LSE-FDB. Hence, it was decided that it was less risky and fairer to crawl the historical price data every week from ADVFN which is provided through a paid subscription. In addition to live prices, LSE-FDB's paid subscription also provides members with access to the premium discussion section. Only the publicly available comments on LSE-FDB are taken into account, considering it is the most influential piece of information for non-paid members as well as investors who perform an internet search. Unlike the premium discussion section (private comments), these publicly available comments will appear as results on the public online search engines such as Google (Google, 2017) and Bing (Bing, 2017) when keywords are searched for. Similar to ADVFN, financial diary (i.e. financial details such as company's turnover, profit, market capital and so on) are also available on LSE-FDB. These financial data are the same across both FDBs.

2.3.3 Interactive Investor

Interactive Investor (III, 2017) was established in 1995. It is also one of the leading investment and trading FDBs in the UK. III has stated that there are over 600,000 messages posted each year and over 1.5 million unique visitors which help to shape its online community. Like ADVFN and LSE-FDB, III not only provides an FDB for investors to exchange financial knowledge but also other information such as prices, news and fundamentals. However, publicly available artefacts on III are fewer than on ADVFN and LSE-FDB. Figure 2.3 demonstrates a sample of how a ticker symbol such as LLOY would appear on the III website and can be navigated around to access other artefacts such as news and discussions.

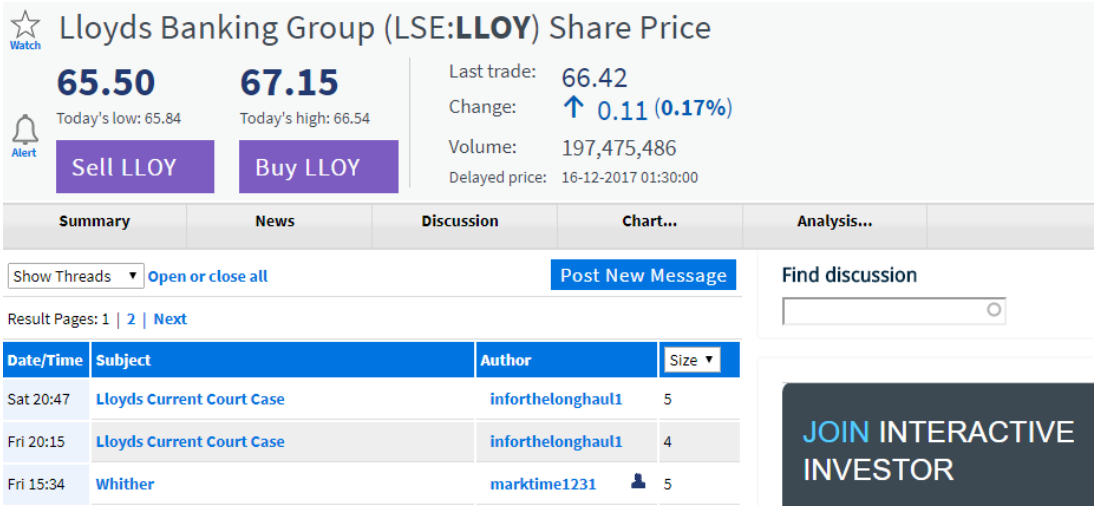

Figure 2.3 Lloyds (LLOY) on III

III also allows investors to subscribe to its monthly paid subscription to stream live prices and gain access to other features such as "Level 2" data which appears to be the stock information at a more in-depth level. "Level 2" data can only be explored by subscription.

All three well established share price based FDBs allow investors to hold discussions while trading on their platforms. On all three FDBs, the terms and conditions (T&C) for using the FDBs were stated. Snippets of the warnings are presented as follow:

> *"…ADVFN does not monitor, approve, endorse or exert editorial control over information posted by users and does not therefore accept responsibility for or make any warranties in connection with or recommend that you or any third party rely on such information."* – ADVFN, 2017.
>
> *"…The contents of all 'Chat' messages should not be construed as advice and represent the opinions of the authors, not those of London South East Limited, or its affiliates."* – LSE-FDB, 2017.
>
> *"…The content of the messages posted represents the opinions of the author, and does not represent the opinions of Interactive Investor Trading Limited or its affiliates and has not been approved or issued by Interactive Investor*

Trading Limited. You should be aware that the other participants of the above discussion group are strangers to you and may make statements which may be misleading, deceptive or wrong." – III, 2017.

Although T&Cs have been made available to the public readers, not everyone follows the rules. Some do it intentionally and some unintentionally. Unlike ordinary online discussion boards, ADVFN, LSE-FDB and III appeared to have not explicitly announced any forum moderator usernames. It is unsurprising that there is no one proactively moderating the FDBs until someone reports a post, considering how time consuming it is to monitor the massive amount of posted comments made in a day. Generally, instead of moderating members' sentiments and opinions, a forum moderator's task is to monitor any spam content, advertisements, forum etiquettes and other generic rules. This has led fraudsters to take FDBs for granted over time to conduct P&D scheme.

2.3.4 Identified Semantically Understandable Artefacts

Table 2.1 represents the financial information available on the three FDBs introduced above for each company listed on LSE.

Table 2.1 Available Financial Information of Each Listed Company

	Description	ADVFN	LSE	III
Share name	The name of the listed company	√	√	√
Share symbol	The unique abbreviated name of the listed company	√	√	√
Market	The stock exchange the share listed on	√	√	
Type	Type of the share	√		
Share ISIN	International Security Identification Number of the share	√	√	√
Share description	Describe the type of share	√		

Price change	How much has the price change on the last day	√	√	√
% change	How many percentages have the price change on the last day	√	√	√
Share price	Current and latest share price	√	√	√
Bid price	Nearest price that the traders want to buy the share at	√	√	√
Ask price	Nearest price that the traders want to sell the share at	√	√	√
High price	Highest price of the day a share reached	√	√	√
Low price	Lowest price of the day a share reached	√	√	√
Open price	The opening price of the day	√	√	√
Close price	The closing price of the day		√	√
Shares traded	The volume of share being traded on the day	√	√	√
Last trade	The latest time of an LLOY share being traded	√	√	√
Industry Sector	Which sector the listed company falls into	√	√	√
Turnover (m)	The liquidity of a company	√		
Profit (m)	The profit made by a company	√	√	
EPS - Basic	The value of one unit of the company's share	√	√	
PE ratio	Market price divided by earnings per share (EPS)	√	√	√
Market Cap (m)	Total market capital (in million)	√	√	√
Related Shares	Shares that are in the similar sector	√	√	
News	Fundamental news related to a listed company	√	√	√
Regulatory News	Official regulatory news of a company announced by RNS (a service provided by LSE) as a form of communication with the professional investors		√	
Share Chat	The discussions of the specific share	√	√	√
Dividends	The dividends distribution historical data	√	√	√

Director Deals	Information of the buy and sell deals by company directors		√	
Broker Ratings	The ratings given by brokers for a company		√	

The extended financial information presented in Table 2.1 is summarised and categorised into Table 2.2. Table 2.2 presents the semantically understandable artefacts identified on the three FDBs. Artefacts such as ticker symbols, prices and financial diary remain the same across all FDBs as these artefacts are extracted directly from the stock exchange market. Artefacts like comments are not identical across the FDBs in terms of the content by different authors, unless there are authors who intentionally write the identical comments on all FDBs, which did not happen in the collected data.

Table 2.2 Identified Semantically Understandable Artefacts

Artefact	ADVFN	LSE	III
Ticker Symbols	√	√	√
Prices	√	√	√
Comments	√	√	√
Director Deals		√	
Broker Ratings		√	
Financial Diary	√	√	
Fundamentals	√	√	
Recent Share Trades	√	√	
Regulatory News		√	

2.4 Stock Market Regulatory Agencies

If a stock market is left uncontrolled and unchecked, it will become a hotbed for fraudsters to manipulate the market and deceive other investors. This is where the stock market regulatory agencies are formed, to prevent financial disasters. With such regulatory agencies, investors are supposed to feel more confident investing and trading in the stock market. However, the scope of the regulators' governance may be too broad, especially on the free internet with currently over 1.3 billion active

websites. Share price based FDBs are part of these websites where they lack surveillances. If share price based FDBs are not properly moderated with maximum efforts for potentially illegal activities like P&D crime, it will become the hub for fraudsters to commit financial crimes, thus, the development of FDBM. Section 2.4.1 describes one of the most preeminent stock market regulators, namely the Security Exchange and Commission (SEC) in the United States. Section 2.4.2 describes the SEC equivalent in the UK, namely Financial Conduct Authority (FCA).

2.4.1 United States of America

The Security Exchange and Commission (SEC, 2017) was formed in 1934. SEC is a securities regulatory agency that is independent of United States (US) federal government. Its mission is to protect investors and regulate the securities market. SEC believes as more and more first timers enter the stock market there is a need for a market regulation. SEC believes the only way for investors to really protect themselves while investing is to do their own research and ask questions. Investors can easily lookup on the internet for regulatory press releases, forum discussions, scam alerts and news before making any investment decisions. Besides, ask questions and seek advice from a financial professional can also reduce the risk of becoming P&D crime victims.

2.4.2 United Kingdom

In the UK, the Financial Services Authority (FSA, 2013) was the equivalent of SEC in the US prior to April 2013. It was then separated into two regulatory authorities, namely Financial Conduct Authority (FCA, 2017) and Prudential Regulatory Authority (PRA, 2017) (i.e. PRA ensures financial firms such as banks to hold enough funds and practice appropriate financial risk controls). FCA then took over the responsibilities of FSA and exercised the powers. Hence, FSA was replaced by FCA in April 2013. FCA's missions are to protect investors, stabilise the UK financial system and reduce financial crimes. FCA can take actions in relation to market abuse by imposing penalties against relevant firms or personnel. On 3rd of July 2016, the Market Abuse

Regulation (MAR) took effect across the European Union (EU). MAR replaces the existing Market Abuse Directive (MAD) and is said to have strengthened the previous rules in MAD.

The key difference of the existing MAD and the new MAR in terms of market manipulation is that the trading firms will now need to develop a more robust, systematic and automated surveillance process to perform surveillances towards all the trades and orders, instead of just a sample of the trades which followed the existing MAD. A stricter rule has also applied in MAR where the firms are also obliged to report the suspicious transactions and orders (STORs) to FCA after implementing new surveillance process. Figure 2.4 shows the increment of suspicious market manipulation reports through STORs from the year 2007 to 2016.

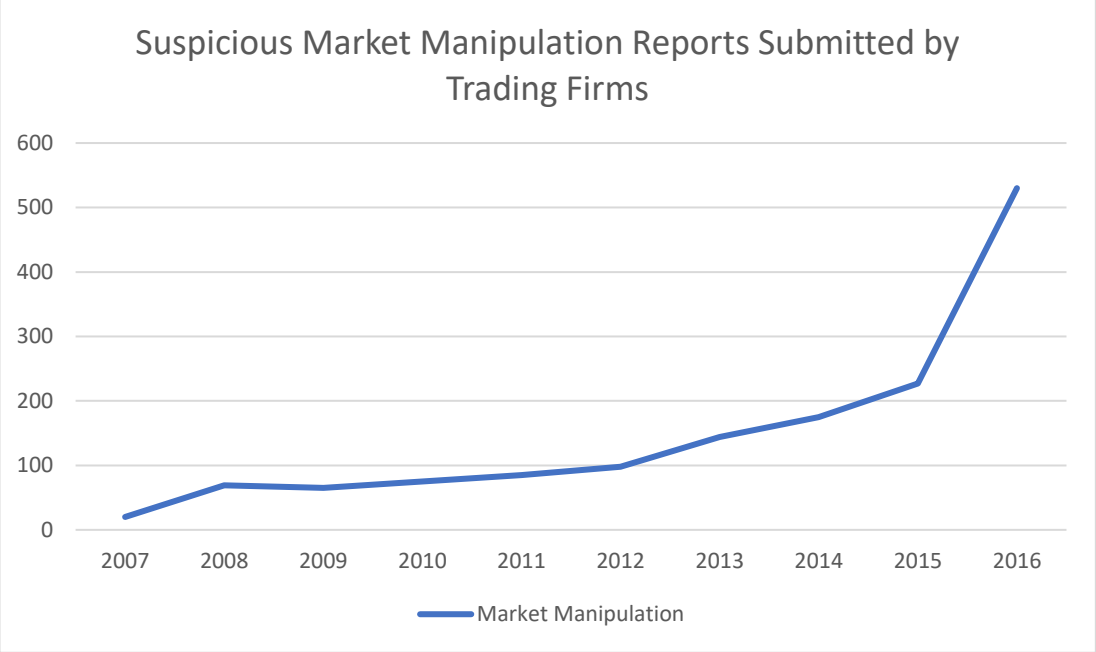

Figure 2.4 Suspicious Market Manipulation Reports Submitted by Trading Firms (FCA, 2017)

FDBs are a popular mode of communication among investors (Halifax UK, 2017). According to MAR's Section 1.8 - Market Abuse, a market manipulation example in relation to FDBs was provided:

"*…a person posts information on an Internet bulletin board or chat room which contains false or misleading statements about the takeover of a company whose shares are qualifying investments and the person knows that the information is false or misleading.*"

With the development of FDBM, it can possibly contribute to the regulatory agencies in the detection and prosecution of potential criminal activities in relation to FDBs without expensive efforts in terms of time and manual human efforts.

2.5 Financial Crimes on Share Price Based FDBs

Traditionally, word of mouth and "boiler rooms" used to be the major ways to commit a financial crime like P&D. According to the SEC (SEC, 2017), "boiler rooms" involve salespersons cold calling victims and promoting penny stocks in a very positive and convincing tone. Cold calling is a method where the salespersons make unsolicited calls to potential buyers in an attempt to promote and sell products. These salespersons are under high pressure all the time, to achieve successful sales; hence, the name "boiler rooms" (Investopedia, 2017). This usually forces the victims to proceed with payment because of the high return promises claimed by the fraudsters. But in fact, these so-called high return investments are actually worthless, non-tradable or non-existent (Financial Times, 2010). Hence, the victims are deceived and lose their money entirely.

As time goes by, with emerging internet trends such as the incremental use of emails, P&D fraudsters have adopted an email method for committing a P&D crime by sending spam emails to promote "good stocks". When free and paid service providers like Google and Symantec improved their spam email filter and detection algorithms (Wired.com, 2015; Symantec, 2017), fraudsters followed the latest internet trends once again to commit financial crimes. This time around the fraudsters began to use FDBs, social media and chatrooms (SEC, 2015) in addition to the old methods that were still not fully eliminated.

With the existence of stock exchanges, market manipulation crimes like P&D are an ongoing issue. However, the number of such crimes supposedly can be reduced with best efforts and useful tools. There are P&D financial crimes being committed from time to time and through different channels. Generally, the SEC and FCA have been enforcing laws and regulating the stock market in the US and UK respectively. But, there is no clear evidence to show that these authorities are actively monitoring share price based FDBs, likely due to insufficient manpower or tools that are specifically developed for monitoring posted comments on FDBs.

Unfortunately, since FDBs are not proactively monitored by human moderators and relevant authorities like the FCA, fraudsters are very likely to be benefited by it. Even if the FDBs are monitored by human moderators or relevant authorities, with such a huge volume of comments posted on different FDBs and for more than 2,000 ticker symbols, it is unrealistic for human moderators to read through all the comments on a daily basis, not to mention the fact that, like any other websites on the Internet, FDBs are meant to be online 24 hours a day. In such situation, there is a need to develop a system to support the monitoring and detection of potentially illegal P&D activities on FDBs.

There have been several popular and significant FDB related P&D financial crimes in the past, which are highlighted as follow:

- 15-year-old Jonathan Lebed was the first minor to be involved in a stock market fraud in 2000 (Lewis, 2001). Lebed earned a total revenue of US$800,000 by pumping the share price through the Yahoo! Finance Message Board over half a year and was charged by the SEC (Lewis, 2001; Reim, 2001; Cybenko et al., 2002).
- In 2000, two people were charged for pumping the price of a share by 10,000% by posting on the Raging Bull message board and then dumping millions of shares with a profit of at least US$5 million (Reim, 2001).
- In 2009, eight participants were charged by the SEC for being involved in penny stock manipulation (SEC, 2009). These wrongdoers met each other through InvestorsHub (now owned by ADVFN), a popular penny stock

message board. This was followed by them carrying out a P&D scheme throughout the year of 2006 and 2007.

2.6 Existing Pump and Dump Related Research

2.6.1 Emails

Siering (2013) conducted an event study alongside a sentiment analysis on a total of 1,299 suspicious stock recommendation newsletter emails that were manually obtained by the author from the "Newsletter Hub" section of the website HotStocked.com (HotStocked, 2017). When selecting these newsletter emails, the author followed the guidelines proposed by the SEC. The guidelines suggested that stocks that are normally involved in P&D are usually from poorly-regulated markets like the FTSE AIM All Share. Also, such emails often lure recipients into buying certain stocks by urging readers to buy the stocks, spreading misleading and exaggerated statements about how well the stocks will perform. In the 1,299 suspicious newsletter emails, a total of 221 stocks were recommended in 252 newsletter emails. According to the author's findings, on average, 252 newsletter emails were sent in five different emails by two P&D campaign promoters during a period of about two days. Based on the dictionary-based sentiment analysis, the author also found that the positive words used in the emails are positively related to the abnormal stock returns. The author concluded that the web and social media still play a significant role in enhancing this type of financial crime despite the efforts by relevant authorities to fight it.

Apart from the newsletter emails which have now been strictly filtered by services such as Google (Google, 2017) and Symantec (Symantec, 2017), newer tactics such as social media and discussion boards were adopted for performing P&D schemes by luring inexperienced investors into buying the so-called recommended stock mainly because these channels allow more freedom of speech.

2.6.2 Social Media

Social media is one of the latest popular methods used by fraudsters to manipulate the market. Twitter is especially popular as it allows the use of “hashtag” (e.g. #LLOY) and “cashtag” (e.g. $LLOY) in each tweet alongside text, links, photos, animated photos or videos (Twitter, 2017).

Renault (2014) and Renault (2017) analysed millions of messages (tweets) on Twitter using network theory and the results showed that there was a spike in the number of posted messages related to securities traded in over-the-counter (OTC) markets on the same day when the prices spiked. It was then followed by a sharp price reversal in the following days. This is consistent with the behaviours of P&D crime.

Wolfram (2010) conducted research using Natural Language Processing (NLP) techniques with a predictive machine learning approach, to examine if the fast-growing social network Twitter can be used to predict share prices. The author selected several stocks from the NASDAQ stock exchange and the intraday price figures were collected for a period of two weeks. Intraday price figures also mean per minute price figures, which are also used in this research. According to Wolfram, most of the prediction models utilise End of the Day price figures rather than intraday price figures. The results were varied across different stock selections, but the author was able to achieve “profits” in some instances.

2.6.3 Financial Discussion Boards

Phillippsohn (2001), a leading authority on fraud in the UK whose his firm has an extensive international network of experienced lawyers and financial crime investigators, has reviewed several types of financial crimes on the Internet including P&D financial crime cases. One of the P&D crimes that the publication’s author discussed was related to a coffee trading company named Coburg PLC based in the UK. The share price of Coburg PLC doubled after a fraudster went on an FDB to spread false news. The share price was also “dumped” on the same day. In another P&D case reviewed by Phillippsohn (2001), eConnect was charged by the SEC in the US for issuing misleading and untrue press releases on an FDB. The false press release

claimed that the company had a unique license arrangement with Palm Inc to permit cash transactions over the Internet. eConnect's share price went from US$1.39 to US$21.88 in just one week.

Delort et al. (2011) introduced a novel classification technique for a classifier training to automate moderation tasks on Online Discussion Sites (ODSs). A partially labelled corpus, i.e. FDB comments taken from HotCopper (HotCopper, 2017, a popular Australian FDB) is used for training purposes and then attempts to moderate the inappropriate content on ODSs using the technique. The authors collected a total of 1.14 million comments involving 1,825 companies listed on the Australian Stock Exchange (ASX) from January 2008 through to December 2008. Of these 1.14 million comments, 14,139 comments were labelled. These comments that were labelled by a human moderator included labels such as ramping, insider trading, profanity, copyright, sexist, racist, flaming, spamming and other generic forum breaches. The authors chose a partially labelled corpus instead of a fully labelled corpus because a partially labelled corpus is much easier to produce as it is divided into the unlabelled comments and the labelled ("bad", as a single class) comments for classification purposes. The authors implemented and tested their technique and 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 a partially labelled corpus. According to the authors, the misclassification errors remain too significant. Besides, the research only takes comments into account and no prices are involved during the classification of comments.

A system named the Financial Discussions Detection System (FDDS) was proposed by Knott and Owda (2012) 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 match the downloaded comments against the potentially illegal keywords created in earlier steps. However, looking only at the comments during the detection process appears to be insufficient as it does not take share prices into account. Thus, this research introduces the novel methodologies

(forward analysis and backward analysis) in an attempt to reduce false positives by integrating share prices in the detection process.

Leung and Ton (2015) examined whether 2.5 million messages posted on the largest FDB in Australia, namely HotCopper (HotCopper, 2017), from January 2003 through to December 2008, had an impact on the ASX market. These 2.5 million messages belonged to all the companies listed on the ASX. There were more than 2,000 listed companies. The results show that the FDB messages had impacts on the small capitalisation stocks but did not affect the large stocks. The authors concluded that the FDB comment posting activities are positively correlated with the trading volume for these small stocks in a highly regulated ASX market.

Alić (2015) 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 prototype system is capable of collecting both unstructured and structured data of selected companies when used by users. The models take into account attributes such as market segment, market capitalisation, trading volume, the age of the company and so on. Subsequently, attribute scales ranging from very low to very high are used for aggregation to determine whether there are suspicious market manipulation activities such as P&D.

Other researchers (Antweiler and Frank, 2004; Cook and Lu, 2009; Bettman et al., 2011) have also found relationships between FDB comments and market performance. Through conducting statistical analysis, the authors reported that the FDB comments could be manipulative and affect share prices.

2.7 Chapter Summary

This chapter has introduced the stock exchange in the UK, namely the London Stock Exchange (LSE, 2017) and its regulatory Financial Conduct Authority (FSA, 2017) as well as the equivalent preeminent authority in the US, namely the Security and Exchange Commission (SEC, 2017). These authorities are responsible for protecting investors from becoming victims of financial crimes.

Three share price based FDBs (i.e. ADVFN, London South East and Interactive Investors) that were created based on LSE market data have been comprehensively described. These FDBs are the most popular share price based FDBs in the UK; they provide the same factual LSE market data except in terms of their discussion sections. These FDBs allow investors to participate in discussions while trading. However, not all investors use FDBs lawfully. Financial crimes like P&D are likely to happen on unmonitored and unregulated FDBs.

There were attempts to moderate the FDB comments by Delort et al. (2011) and Knott and Owda (2012). However, misclassification remains high in Delort et al.'s (2011) attempt to moderate. Both research attempts did not take share price into account while moderating FDB comments to reduce false positives. Besides, the existing research (Antweiler and Frank, 2004; Cook and Lu, 2009; Bettman et al., 2011; Leung and Ton, 2015) has also found that FDB activities can impact on the stock markets' performance.

Therefore, to address the challenges of FDBs having to be proactively monitored and resolve gaps identified in the existing research, share prices are taken into account when flagging and detecting potentially illegal FDB comments, in this study. This leads to the introduction of two novel built-in methodologies (namely, forward analysis and backward analysis) and the integration of Semantic Textual Similarity (STS) (Chapter 7) for flagging potentially illegal comments and resolving false positives during the comments flagging process.

To devise and implement these methodologies, Chapter 3 conducts an investigation into the types of data, the field of Information Extraction (IE) and STS. These fields are used for the development of novel methodologies used within the template-based Financial Discussion Boards Miner (FDBM) prototype system proposed in the forthcoming chapters.

Chapter 3. Information Extraction and Semantic Textual Similarity

3.1 Introduction

Previously in Chapter 2, three share price based Financial Discussion Boards (FDBs) in the UK, namely ADVFN (ADVFN, 2017), London South East (LSE-FDB, 2017) and Interactive Investors (III, 2017), have been identified and discussed. In order to develop a template-based Information Extraction (IE) prototype system, namely the Financial Discussion Boards Miner (FDBM), to extract the artefacts from these FDBs the area of IE will be studied and reviewed in this chapter.

This chapter first defines and introduces the three types of data structures on the Internet. It goes on to introduce the two fundamental classes of IE, namely a knowledge engineering approach and an automatic learning approach. The template-based IE prototype system, namely FDBM, introduced in this thesis, is based on the knowledge engineering approach.

FDBM automatically extracts information from an unstructured or semi-structured data source (such as FDB comments, FDB Really Simple Syndication (RSS) feeds and share prices) into a structured data format (i.e. FDB dataset). The IE prototype system in this research will be used to 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.

Section 3.4 includes an overview of Semantic Textual Similarity (STS) algorithms, including a description of STASIS – an original and well cited algorithm for defining similarity between short texts. Applications of STS measures are also discussed to justify its suitability for the application of FDB textual comments. Lastly, Section 3.5 summarises the chapter.

3.2 Unstructured, Semi-Structured and Structured Data

Sir Tim Berners-Lee, a British computer scientist, invented the World Wide Web (WWW) in 1989 (World Wide Web Foundation, 2017). In 1990, Tim successfully wrote three ultimate technologies which remain the foundation of the web today. These technologies are the HTML (HyperText Markup Language), URI (Uniform Resource Identifier) and HTTP (Hypertext Transfer Protocol). HTML is the primary formatting language for websites. URI is commonly called the URL (Universal Resource Locator) nowadays. URI is a unique address to identify the resource on the web.

A complete URL looks like this: http://www.webfoundation.org, whereby these days HTTPS (HTTP Secure) is more favourably used than the basic HTTP due to a rule imposed by Google (Google, 2017). A website that does not use HTTPS will not be ranked higher in the search results. HTTPS is not only great for a website's reputation nowadays; the primary purpose was initially to encrypt the transmission of sensitive data between the backend server and the frontend website input by website users. The data encryption process (i.e. the process of transforming information into an unreadable format) prevents hackers from prying into the plaintext data while it is being transferred.

With the rapid growth of publicly available data on the Internet in the past two decades, data have developed different structures: structured data, semi-structured data and unstructured data. Structured data are the data that can fit into relational databases in rows and columns, are easy to manage and use for performing search queries. HTML, XML (Extensible Markup Language) and CSV are the usual forms of semi-structured data. Semi-structured data are human-readable and machine-readable. The useful data reside in specific properties in HTML and XML files, which can be processed by a computer and stored in the relational databases. The same goes for CSV; the data in CSV can be processed and stored in the databases. Unstructured data covers most data formats such as discussion board posts, social media data, email content, customer service live chat transcripts, multimedia content and more. Unstructured data represent the most complex data structure to be processed as they cannot be neatly fitted into the databases.

3.2.1 RSS Feeds

The comment artefact data in this research are collected through RSS feeds on the FDBs. RSS stands for Really Simple Syndication (Software Garden Inc., 2004). RSS, also known as news feed, is a technology that allows readers to keep track of their favourite websites with the latest posts or news. Bookmarking websites is deemed complicated (Problogger.net, 2017) if the posts that a user wants to keep track of keep increasing in number. Also, the user has to do the work by bookmarking the posts manually which is time consuming. Often, when a site is RSS Feed activated, there will be a logo beside the URL in the URL bar. Figure 3.1 below describes the look of a URL bar if RSS feed is available.

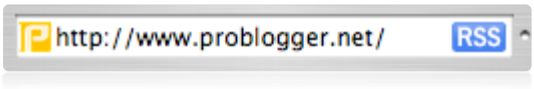

Figure 3.1 RSS Feed Icon on URL Bar (Problogger.net, 2017)

RSS feed is often read by users in a web browser itself, integrated into email programs or read by standalone software on a computer. Figure 3.2 below shows how the RSS Feed XML files, website and computer are connected. A user uses the web browser on his personal computer to visit website 1 and also website 2. Then, the RSS Feed Aggregator will monitor the RSS feed for both websites the user wants simultaneously.

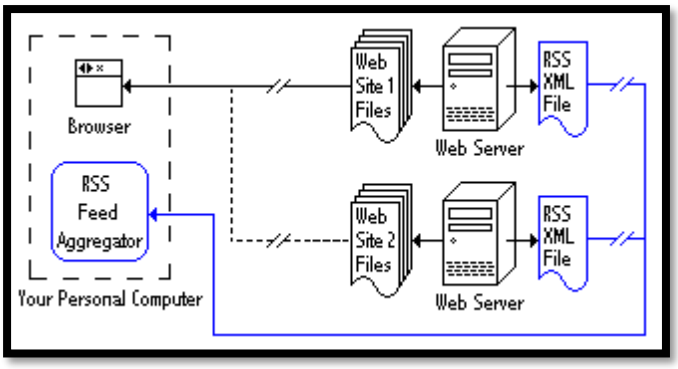

Figure 3.2 RSS Feed Structure (Software Garden Inc., 2004)

RSS is not only useful to keep track of forum topics, but it can also be helpful for other purposes (Software Garden Inc., 2004) such as for shoppers who are keeping track of new products on shopping websites or for monitoring weather alerts.

LSE-FDB used to provide RSS feed, but not anymore. III still provides it. ADVFN does not have it.

3.3 Information Extraction (IE)

Masterson and Kushmerick (2003) defined IE as the process of extracting information automatically into a structured data format from unstructured or semi-structured data sources. It was suggested by Soderland (1999) that there is a need for systems that extract information automatically from text data. IE is not Information Retrieval (IR) (Cunningham, 2004). IE is said to be a subfield of the broader field of IR. Although IE has not received as much attention as IR, IE techniques have been explored and employed in many models and systems over the past two decades. There is a significant difference between IE and IR. IE systems are knowledge-intensive as these systems extract only snippets of information that will predefine templates (fixed format) or databases, which represent useful and relevant information about the domain, and present to the users for actions such as decision making. IR, on the other hand, finds data in the form of ranked document lists which do not display any detailed information from the document and present the located documents to the users based on the given query (Piskorski and Yangarber, 2013). Table 3.1 displays an

example of the extracted information, from a snippet of news, that is fitted into a predefined template.

> "*The Securities and Exchange Commission today charged a New Jersey man (Samuel DelPresto) and his company with illicitly pocketing $13 million from an elaborate pump-and-dump scheme*." (SEC, 2015)

Table 3.1 An Example of Extracted Information from SEC News

Regulator	Security Exchange Commission
Financial Crime	Pump-and-dump
Amount	$13 million
Offender	Samuel DelPresto
State	New Jersey

IE is divided into two fundamental classes i.e. a knowledge engineering (KE) approach and an automatic training approach (Appelt and Israel, 1999). The KE approach is also called the rule-based approach since it requires rules to be developed by human expertise. Rules are in the form of "IF some condition THEN some action" (Ireson-Paine, 1996). Rule-based IE systems require manual effort for rule writing. However, they are easy to maintain and comprehend. Errors can also be traced and fixed easily. On the other hand, the automatic training approach, also known as the machine learning approach, is quite the opposite of the rule-based approach. It does not require human expertise to write rules. Instead, it requires only someone who is familiar enough with the domain and able to train the corpus using machine learning algorithms. Although the machine learning approach requires less manual effort, the approach requires pre-labelled data and retraining for adaptation (Chiticariu et al., 2013).

According to Chiticariu et al. (2013), the rule-based approach is typically ignored by the research community due to the rule-writing labour cost, which directly motivates the researchers to write tools using the automatic training approach that are deemed better. Given that the research community has shifted its focus towards the

automatic training approach for the past decade, it would be reasonable to assume that the automatic training approach would be adopted in the commercial world. However, the opposite turns out to be true. In fact, the rule-based approach is mostly favoured in the commercial market even by larger vendors such as IBM for text analysis systems (IBM, 2017) and Microsoft for its enterprise search platform (Microsoft, 2017).

In this research, the rule-based approach was adopted due to the nature of the system needing to detect potentially illegal comments on FDBs.

3.4 An Overview of Semantic Similarity Measures

3.4.1 Applications of Semantic Similarity Measures

STS plays a significant role in text related research and the research in this area continues to emerge. STS measures the similarity between two sentences generally ranging between 10 and 25 words (Chandran et al., 2015) which then returns a similarity score of between 0 and 1, where 1 means extremely high similarity. STS has been applied in different areas such as information retrieval from biomedical ontologies (Couto and Pinto, 2013), joke detection in the Japanese language (Rzepka et al., 2015), a short answer grading system (Sultan et al., 2016), conversational agents (Sandbank et al., 2017), a tutoring system (Aljameel et al., 2017) and events detection through social media (Crockett et al., 2017).

There have been a number of STS measures introduced previously, which can be categorised text into string-based, corpus-based and knowledge-based measures (Gomaa and Fahmy, 2013). String-based similarity measures the similarity between the strings of two words. For example, “kitten” and “mitten” can be considered similar. Two of the most popular STS measures are Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997; Landauer et al., 1998) and Semantic Similarity based on Semantic Nets and Corpus Statistics (STASIS) (Li et al., 2006). LSA is a corpus-based similarity measure whereas STASIS is a knowledge-based similarity measure.

Corpus-based similarity, like LSA, is a similarity measure that measures the similarity between words according to the statistical relationship of vocabulary gained from the large corpus (Lee et al., 2014). When using LSA, a matrix of the word by paragraph (row represents word and column represents paragraph) is formed from a large piece of text. Take, for example, an article. LSA uses a mathematical technique called Singular Value Decomposition (SVD) to find the semantic similarity of texts. SVD removes the smaller singular values so that the dimension of the word by paragraph matrix is reduced. The vectors of the words are transformed into a reduced dimensional space while the similarity between two texts is obtained by calculating the vectors in the reduced dimension (Foltz et al., 1998). LSA does not take into account function words (such as “at”, “the” and “that”) and word order (i.e. syntactic information) while calculating for semantic similarities. A grammatically correct sentence is not required during the computation of similarities when using LSA.

Knowledge-based similarity, like STASIS, measures the similarity between words using the information derived from a large lexical database in English such as Wordnet (Miller, 1995). Wordnet can be obtained through the Natural Language Toolkit (NLTK) (Bird, 2006). NLTK is an open source tool that includes more than 50 lexical and corpora sources including Wordnet. Wordnet is like a thesaurus, where the relationships between words mainly relate to synonymity (Fellbaum, 2006). Adjectives, adverbs, nouns and verbs are all categorised into groups of cognitive synonyms, i.e. synsets. STASIS compares two texts by using short vectors. STASIS takes function words and word order into account when calculating for similarities. The dimension vector in STASIS varies according with the sentence pair; hence, according to the author (Li et al., 2006), their algorithm is a lot more computationally efficient than an algorithm like LSA. Like LSA, STASIS does not require grammatically correct sentences when computing similarities.

Since STASIS takes both function words and word order into account and does not need grammatically correct sentences, it is more suited for use in Experiment 2 of Chapter 7. Word order is important when it comes to texts like FDB comments because it is a form of communication among investors on FDBs. They do not just post words without conveying information or meaning.

3.5 Chapter Summary

This chapter has reviewed the three types of data structures that can be collected from the Internet. It then introduced the two fundamental classes of Information Extraction (IE), namely the knowledge engineering approach and the automatic learning approach. The knowledge engineering approach, which also known as the rule-based approach is the approach used by the template-based IE prototype system, namely FDBs Miner (FDBM) in this thesis.

The next chapter introduces new methodologies and then presents an architecture for financial crime detection from share price based FDBs.

Chapter 4. A Methodology for Financial Crime Detection on Share Price Based Financial Discussion Boards

4.1 Introduction

This chapter introduces the development of a set of novel methodologies for the detection of potentially illegal activities on FDBs. Section 4.2 describes the identification of share price based FDBs in the UK and the semantically understandable artefacts on these FDBs. Lastly, Section 4.3 presents each methodology in each phase of the detection of potentially illegal activities on FDBs.

4.2 Identification of Share Price Based FDBs and Semantically Understandable Artefacts

Before the detection of potentially illegal activities can be performed, share price based FDBs based in the UK and semantically understandable artefacts were identified using the following steps:

i. Manual search on the internet for share price based FDBs based in the UK.
ii. A decision on which London Stock Exchange (LSE, 2017) indices to use for this research.
iii. A randomly picked number of stock tickers from the selected UK indices.
iv. A study of the stock tickers' webpages and identification of the semantically understandable artefacts on all the identified FDBs in step (i).

The identified share price based FDBs in the UK are ADVFN (ADVFN, 2017), London South East (LSE-FDB, 2017) and Interactive Investors (III, 2017) which have been comprehensively discussed in Chapter 2. As for the LSE indices, the FTSE-100 and FTSE AIM All-Share indices have been chosen because this allows the observations of whether Pump and Dump (P&D) does happen more in the FTSE-100, or FTSE AIM All-Share or both equally. As mentioned in earlier chapters, P&D usually happens around penny stocks (Leung and Ton, 2015) and most of these penny stocks are listed under

the FTSE AIM All-Share index, hence, this choice. The exact number of total listed companies, unfortunately, was not identified at the beginning of this research in 2014. However, according to a recent list of listed companies on the London Stock Exchange (LSE, 2017) website, there was a total of 2,037 companies trading shares on the market as of May 2017. The FTSE-100 and FTSE AIM All-Share indices both have a total of 941 listed companies at the time of data collection in September 2014, which accounted for nearly half of the total companies.

All possible semantically understandable artefacts were also identified for all the identified share price based FDBs. This factor was also discussed and presented in Chapter 2. However, due to the scope of this research, the complexity of the webpages' format and the number of artefacts on the FDBs, not all artefacts were extracted.

During the examinations of FDBs in Chapter 2, it was found that only ADVFN provides downloadable per minute historical prices (in the form of downloadable CSV files) with the condition that a monthly paid subscription needs to be bought. These downloadable per minute historical prices contain a maximum of two weeks' worth of data. Thus, it was crawled once per week, so that no price data were accidentally left out. Though the LSE-FDB also provides per minute prices through a paid subscription, it is only available in the form of live prices. This means that there would be a need to develop another custom crawler function to extract the per minute live prices from the LSE-FDB. Assuming that this would even be achievable without too much effort, the owner of the LSE-FDB did express his disapproval (via email communication) of scraping the live prices from the LSE-FDB webpages. The owner's disapproval was understandable as crawling per minute of live share prices from raw webpages for a span of 12 weeks would cost the owner in terms of having to pay for premium bandwidth charges because each crawl counts as a visit to the website. Hence, for the share prices artefact, ADVFN is being utilised.

Publicly posted comments by investors are available on all ADVFN, III and LSE-FDB websites. Private comments are available on the LSE-FDB only for paid members, but it was not considered. Public comments are thought to have more influence on the potentially illegal activities on FDBs as they are publicly viewable by non-paid

members and any website visitor. Only the public comments of III and LSE-FDB were collected through RSS feed. ADVFN does not provide RSS feed and due to the complexity of the webpage's format, ADVFN's comments were not considered.

Table 4.1 below summarises the FDB artefacts that were collected from each FDB and SharePrices[1] – a sister website of LSE-FDB (LSE-FDB, 2017) which is no longer active. Director deals and broker ratings were collected for potential future work.

Table 4.1 Collected FDB Artefacts

	SharePrices	ADVFN	III	LSE-FDB
Ticker Symbols	√			
Prices		√		
Comments			√	√
Director Deals				√
Broker Ratings				√

These FDB artefacts were collected for a period of 12 weeks based on the same method adopted by authors such as Delort et al. (2011) and Leung and Ton (2015) (i.e. the data is collected based on a period of time rather than data volume).

4.3 An Overall Methodology for the Detection of Potentially Illegal Activities on FDBs

This section describes the methodology developed for each phase for the detection of potentially illegal activities on FDBs. There is a total of seven phases described in the following sections, which are then implemented in the seven architecture components in Chapter 5.

[1] This website was originally used for obtaining the ticker symbols belonging to both indices FTSE100 and FTSE AIM All-Share in 2014; however, it is no longer active.

4.3.1 Phase 1: Implement a Data Crawler

Development of a data crawler architecture component to extract artefacts identified in Section 4.2.

i. Scrape two webpages (in HTML file format) from SharePrices[1] that contain ticker symbols belonging to the chosen indices – FTSE-100 and FTSE AIM All-Share.
ii. Study the webpage structures and extract only the ticker symbols and company names.
iii. Automatically capture the comments (in XML file format) for these ticker symbols from FDBs for every 10 minutes and store in local folders with timestamps for later use.
iv. Automatically capture the per minute historical share prices (in CSV file format) for these ticker symbols every week from ADVFN via paid subscription. Downloaded share prices are also stored in local folders with timestamps for later use.
v. Collect director deals and broker ratings (both in HTML file format) artefact data for potential future work.

A total of 941 ticker symbols were successfully extracted from the SharePrices website. Comments, share prices, director deals and broker ratings were also captured. This phase is further elaborated and implemented in Section 5.3.

4.3.2 Phase 2: Implement a Data Transformer

Implementation of a data transformer architecture component that is capable of pre-processing and transforming unstructured and semi-structured artefacts data collected in Phase 1 into structured data.

i. Programmatically locate the semantically understandable artefacts data in the indices' HTML files, comments' XML files, share prices' CSV files data crawled in Phase 1.
ii. Extract and transform the revealed data from a semi-structured format such as HTML, XML and CSV into a structured format.

iii. Import the structured data into a database designed in the next phase (Phase 3).

Phase 2 is further described and implemented in Section 5.4.

4.3.3 Phase 3: Devise a Dataset for Storing Data

Phase 3 devises and implements an FDB dataset (FDB-DS) to store the pre-processed structured data extracted in Phase 2.

i. Design the database tables in FDB-DS to accommodate the pre-processed data of ticker symbols, share prices, comments, director deals and broker ratings.
ii. Link all the database tables to the unique ID of the ticker symbols' table. This allows all other artefacts database tables to be linked and related to the ticker symbol.
iii. Decide on the data types for each data column in the database tables. For example, the data type for ticker symbols and company names is "varchar" (i.e. variable character, meaning the column can hold letters and numbers).
iv. Create all database tables in FDB-DS based on the decided table relationships and column data types.

Phase 3 is further elaborated in Section 5.5.

4.3.4 Phase 4: Construct a Pump and Dump Keyword Template

Design of a methodology for the construction of P&D IE keyword template is as follows. This template is used in the process of detecting potentially illegal activities on the identified share price based FDBs.

i. Evaluate sample comments of P&D financial crime presented in existing research (Campbell, 2001; Felton and Kim, 2002; Campbell and Cecez-Kecmanovic, 2011; Sabherwal et al., 2011) (see Section 2.6.3).
ii. Manually select keywords, phrases and short sentences of P&D financial crime from the sample comments.

iii. Look up for similar keywords and phrases on WordWeb (WordWeb Software, 2017), Chamber Dictionary (Chamber Dictionary, 2017) and Thesaurus (Dictionary.com, LLC, 2017).

iv. Form a P&D IE keyword template by adding the keywords, phrases and short sentences into a text file for later use.

v. Validate the list of keywords, phrases and short sentences through a human expert in the relevant field (through email communication).

The formulation of the P&D IE keyword template and the full list of the keywords are presented in Section 5.6.

4.3.5 Phase 5: Devise a Forward Analyser

A novel methodology is devised in Phase 5 for the forward analyser architecture component to perform flagging of potentially illegal activities on FDBs. The flagging process uses a combination of comments and prices from FDB-DS designed in Phase 3 as well as the P&D IE keyword template constructed in Phase 4.

i. Search the keywords, phrases and short sentences in P&D IE keyword template against all the comments in FDB-DS.

ii. Import the result generated in Step (i) (i.e. flagged comments deemed potentially illegal) into a new database table in FDB-DS.

iii. Match each flagged comment with the price of the same ticker symbol and same datetime. If no exact datetime is available, the next available price will be used to match the comments.

iv. Set the matched price of each flagged comment as a base price, calculate whether at any point, the ±2 days' prices exceed the specified thresholds (i.e. 5%, 10% and 15%). Record the outcomes in FDB-DS. Such thresholds can be adjusted by the relevant authorities as needed.

This phase is further elaborated in Section 5.7.

4.3.6 Phase 6: Devise a Backward Analyser

A novel methodology is devised for the backward analyser architecture component to perform moving average calculation, price spike detection on the share price data alone and backward price alert matching to the flagged potentially illegal comments produced in Phase 5.

i. Study various moving average techniques and the use of these techniques.
ii. Decide which moving average technique to use as part of the backward analysis.
iii. Perform moving average calculation on the prices of all ticker symbols.
iv. Depending on the calculation outcomes, label each price with price alerts if its moving average figure exceeds the thresholds of 5%, 10% and 15% of increments.
v. Match and append the price alerts, if any, backward to the flagged comments that share the same or nearest datetime as the prices.

This phase will produce a complete dataset of flagged comments with all the thresholds applied through forward and backward analysis, which can be used for determining which flagged comments should be prioritised for investigation of potentially illegal activities on FDBs. Phase 6 is expanded in Section 5.8.

4.3.7 Phase 7: Implement Semantic Textual Similarity

STS is implemented in Phase 7 to incorporate with forward analyser in Phase 5 with the purpose of reducing false positives during the comment flagging process.

i. Generate a 32-row comment dataset for STS experiments by randomly choosing and sorting 16 flagged comments and 16 non-flagged comments programmatically.
ii. A human expert is asked to determine the flagged and non-flagged comments.
iii. Human expert's answers are compared to the original outcomes (i.e. 16 flagged and 16 non-flagged comments that are detected during forward analysis process by FDBM prototype system).

iv. Next, using the similarity thresholds of 0.5, 0.55, 0.6, 0.65 and 0.7, compute the similarity score between each comment in the 32-row comment dataset and each keyword, phrase and short sentence from the keyword template constructed in Phase 4.

v. Record all similarity results for discussion purposes.

Phase 7 is further elaborated in Chapter 7.

4.4 Chapter Summary

This chapter introduced a set of novel methodologies for the creation of the FDBM architecture, which has been designed to assist the relevant authorities in detecting potentially illegal FDB comments. It also allows the relevant authorities to focus on investigating the flagged comments that have higher risks or higher potentials of association with illegal activity on the FDBs. Section 4.2 covered the area of the identification of share price based FDBs in the UK and the identification of the semantically understandable artefacts on the FDBs. Section 4.3 provided an overview of a set of novel methodologies for the creation of FDBM architecture and a prototype system which will be discussed in Chapter 5.

Chapter 5. An Architecture for Financial Crime Detection on Share Price Based Financial Discussion Boards

5.1 Introduction

Section 5.2 illustrates an architecture overview of the Financial Discussion Boards Miner (FDBM) prototype system based on the methodologies introduced in the previous chapter. Within this chapter, the seven architecture components of the FDBM are described. Sections 5.3 to 5.8 expand the descriptions of the architecture components presented in Section 5.2 which were implemented with the potentially illegal activities methodologies. The architecture of FDBM is comprised of the data crawler, data transformer, FDB dataset (FDB-DS), Pump and Dump Information Extraction (IE) keyword template, forward analyser, backward analyser and Semantic Textual Similarity (STS). The graphical user interface (GUI) for the prototype system is also presented in Section 5.9. Finally, Section 5.10 summarises the chapter.

5.2 Prototype Architecture Overview

Figure 5.1 illustrates an architecture overview of the prototype system FDBM. The FDBM consists of seven key components. As shown in the overview, there are the data crawler, data transformer, FDB dataset (FDB-DS), IE keyword template, forward analyser, backward analyser and STS.

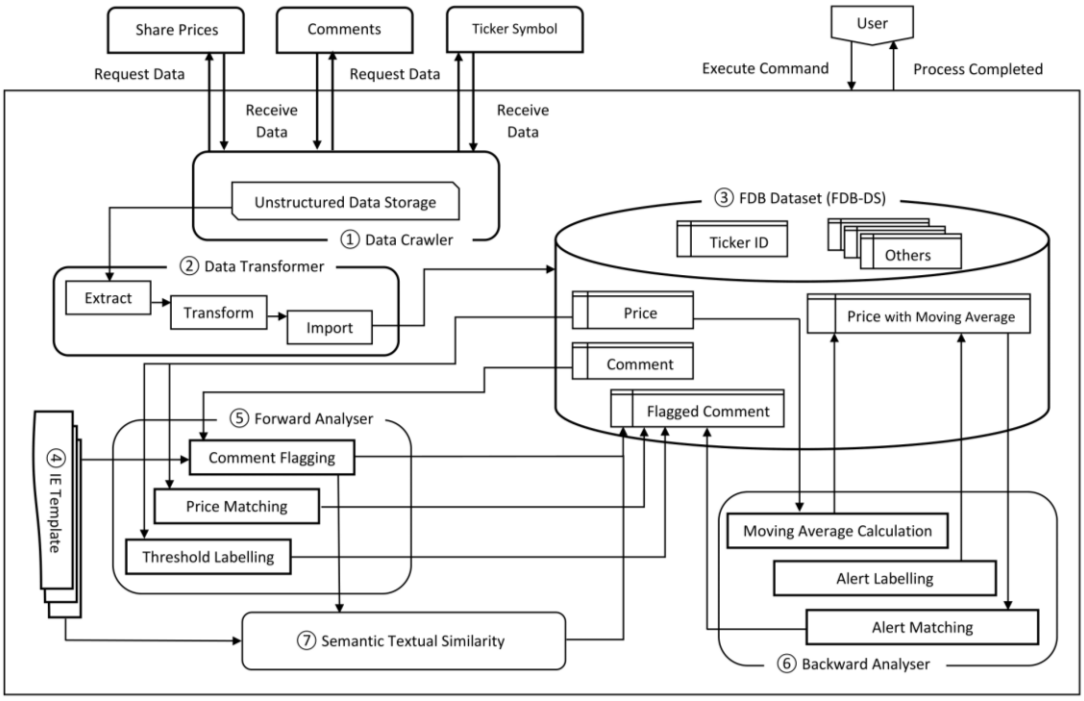

Figure 5.1 Prototype Architecture Overview

Each key component is described as follows:

1. Data Crawler

The data crawler component is implemented to crawl unstructured data such as company names and ticker symbols from SharePrices[2] – a sister website of London South East (LSE, 2017). The ticker symbols were obtained based on the FTSE-100 index and FTSE AIM All-Share index for the London Stock Exchange (LSE, 2017). The ticker symbols' rankings change from time to time depending on the companies' performances. A total of 941 ticker symbols were found in both indices at the time of crawling. The data crawler also crawls the three selected UK share price based FDBs, i.e. ADVFN (ADVFN, 2017), LSE-FDB (LSE-FDB, 2017) and III (III, 2017) for share prices, comments, director deals and broker ratings which belong to all the chosen ticker symbols. These data are crawled

[2] This website was originally used for obtaining the ticker symbols belonging to both indices FTSE100 and FTSE AIM All-Share in 2014; however, it is no longer active.

automatically at different time intervals for 12 weeks from the date of 23rd September 2014 to 22nd December 2014 (see Section 5.3).

2. Data Transformer

Once the data collection process is completed by the data crawler, the data transformer extracts and converts the collected unstructured and semi-structured data in various formats such as HTML, CSV and XML into a structured data format. The data transformer will import all the structured data into FDB-DS once it's designed and prepared for the use of data storage (see Section 5.4).

3. FDB Dataset (FDB-DS)

From the data transformer, the structured data are stored into the FDB-DS in each table as `ticker`, `price`, `comment`, `directordeal` and `brokerrating` respectively. More database tables follow once the forward analysis and backward analysis experiments are performed. FDB-DS becomes the most vital component since the forward analyser and backward analyser will depend on it for conducting experiments (see Section 5.5).

4. IE Template

The P&D IE keyword template is created based on the research methodology proposed in Section 4.3.4. It consists of a series of keywords, phrases and short sentences that were thoroughly researched and validated by a human expert in the relevant field. It is saved locally in the FDBM prototype system in a text file (TXT file format) which can be easily retrieved and modified from time to time. This IE keyword template will be used by the forward and backward analysers to flag potentially illegal comments (see Section 5.6).

5. Forward Analyser

Once the data are properly stored in the FDB-DS and the P&D IE keyword template is validated, the forward analyser incorporates the IE keyword template into the process of flagging potentially illegal FDB comments. The flagged potentially illegal comments are stored in a new database table. Following this, the price figures and price datetime are searched for and matched to the list of flagged comments. Then, ±2 days of prices are calculated against the price of each flagged comment to see if, at any point, the ±2 days of prices exceeds the predefined thresholds (i.e. 5%, 10% and 15%). For any triggered thresholds, the forward analyser will append threshold alerts to the flagged comments (see Section 5.7).

6. Backward Analyser

The backward analyser is responsible for performing the calculation and labelling of any price hikes using a price moving average technique, namely Simple Moving Average (SMA). There are also three thresholds (i.e. 5%, 10% and 15%) for the price hikes. Once all the price hikes for all the tickers are determined through the SMA calculations, the backward analyser then matches the price hike alerts backwards to the flagged comments (see Section 5.8).

7. Semantic Textual Similarity (STS)

The STS component in the FDBM prototype system measures the similarity scores between each FDB comment and each keyword, phrase and short sentence. The similarity scores are used to determine the likelihood of an FDB comment being suspicious. There is a full explanation of this component in Chapter 7.

5.3 Data Crawler

This section describes the role of the data crawler component starting from the collection of ticker symbols (Section 5.3.1), share prices (Section 5.3.2), comments (Section 5.3.3) and other artefact data (Section 5.3.4). Section 5.3.5 presents an overview of the data collection.

5.3.1 Collecting Ticker Symbols

Before collecting artefacts like share prices and comments, the crawler first determines all the participating shares in the FTSE-100 and FTSE AIM All-Share indices by scraping HTML webpages, i.e. http://shareprices.com/ftseaimallshare (LSE-FDB, 2017) and http://shareprices.com/ftse100 (LSE-FDB, 2017) from SharePrices (LSE-FDB, 2017). These HTML webpages are temporarily saved in the local machine for extraction of the ticker symbols and company names later. The listed companies in both indices change over time. However, the process of obtaining these company names and ticker symbols runs only once as the research will only take into account the data for 12 weeks. A total of 941 ticker symbols were extracted and written into a CSV file and imported into the `tickers` table in the FDB-DS.

5.3.2 Collecting Share Prices

Now that all ticker symbols in both indices are identified, the data crawler starts collecting share prices. Per minute share prices are not available publicly and can only be obtained through paid subscription. In this case, ADVDN is chosen for such a purpose and the subscription fee is £15.79 per month. Although this research takes only 12 weeks of data, the paid subscription lasted for about a year due to the testing phase in the development of the FDBM prototype system and data loss in the middle of the data collection process due to hard disk failure. Also, note that ADVFN does not guarantee the availability of all per minute price figures. This means there will be missing figures which will be handled in Section 5.4.3.

To obtain the per minute share prices of all chosen ticker symbols, the data crawler first triggers and opens a new browser window. It then visits http://uk.advfn.com

(ADVFN, 2017) and automatically logs in to a preregistered ADVFN account with an active monthly subscription. Once logged in, the crawler extracts and downloads the prices of each ticker symbol in comma-separated values (CSV) files. These files are saved into dated folders on the local machine for later use. This collection of share prices automatically runs once a week.

5.3.3 Collecting Comments

While the semi-structured price data are being collected, the crawler downloads the comments for all chosen ticker symbols from both the LSE-FDB (LSE-FDB, 2017) and III (III, 2017) in XML file format. These comments are automatically collected every ten minutes to ensure no comments are missed. The crawled comments data are placed in separated and dated folders subsequently. As these data are crawled so frequently, the folder size grows rapidly.

5.3.4 Collecting Other Artefacts Data

Apart from ticker symbols, share prices and comments, the data crawler also captures the director deals and broker ratings which are potentially useful in expanding the research in the near future. Both director deals and broker ratings are captured once a week.

5.3.5 An Overview of the Data Collection

Table 5.1 summarises the frequency with which each capturing process takes place:

Table 5.1 Time Intervals for Capturing Data

Data	Interval
Ticker Symbols	Only once
Share Prices	Once a week
Comments	Every 10 minutes
Director Deals	Once a week
Broker Ratings	Once a week

Once the 12 weeks' worth of artefact data have been successfully captured, the data crawler proceeds to the data pre-processing stage. The crawler was designed to clean up the noise in each semi-structured data file and extract only the meaningful data. This process transforms semi-structured data into structured data. Subsequently, the extracted artefact data will be imported and an FDB dataset (FDB-DS) will be formed which will be described in Section 5.5.

All the acquired data was successfully imported into the database FDB-DS. Table 5.2 shows the total entries of data for each artefact stored in database tables in the FDB-DS:

Table 5.2 Total Database Entries for Data Collected

Artefact	Total Rows
Ticker Symbols	941
Comments	507,970
Prices	28,980,465
Director Deals	11,456
Broker Ratings	6,469

5.4 Data Transformer

This section describes the functionality of the data transformer component and the pre-processing steps for the collected ticker symbols (Section 5.4.1), comments (Section 5.4.2), share prices (Section 5.4.3) and other artefact data (Section 5.4.4).

5.4.1 Pre-processing Collected Ticker Symbols

Previously in Section 5.3.1, the data crawler scraped the HTML webpages for FTSE-100 and FTSE AIM All-Share and stored the webpage files locally. In this section, the data transformer searches for tables in the webpage files and iterates through each row and column. It then extracts the ticker symbol and company names in each row and writes them in a text file. The text file will be used to import the ticker records into the FDB-DS in Section 5.5.

5.4.2 Pre-processing Collected Comments

Previously in Section 5.3.3, the comments for 941 ticker symbols were downloaded from FDBs in XML file format every 10 minutes for 12 weeks and placed into separate folders. Folders were named in the format of "LSE_" + "dd-MM-yyyy_HH-mm-ss" e.g. LSE_02-10-2014_21-10-05, which then makes it easier to convert the XML to CSV in order to import the data into the MySQL database.

By using the designed steps in the data crawler component, XML files were converted to CSV using the following steps:

i. Loop through each of the XML files in each of the folders.
ii. In each XML file, scan for each XML node (which represents each comment).
iii. Further scanning into each XML node, there were XML items such as comment title, comment author, comment and comment date and time.
iv. In order to avoid conflict with the CSV file format, certain visible and invisible characters in the content of XML items were removed as part of the pre-processing steps. These components included: “,” (comma), “\n” (new line), “\t” (tab), “\r” (return – like new line) and “
” (break a line).
v. Write all the pre-processed and cleaned up comments into separate CSV files to prevent an overlarge single CSV file.

RSS feeds contain a fixed number of the latest comments. For the RSS feeds XML file format for the comments being collected, there were duplicated comments which were written into the CSV files. At this stage, the data crawler component is also responsible for removing the duplicates. This process takes a long time as there were a lot of comments to process. The CSV files then went through special character removal such as “, = "" '' * + - %” and each comment row in the CSV files was parsed into the `comment` database table in the FDB-DS.

5.4.3 Pre-processing Collected Share Prices

All the downloaded price file names have a fixed format like “L-<ticker symbol>_<date>_intra.csv”. Remove the “L-” and “_<date>_intra” so that the file names are simplified as “<ticker symbol>.csv”.

Since the experiments throughout the research rely on the ticker ID instead of ticker symbols or company names, this step replaces all the “<ticker symbol>” with “<ticker ID>”, so the file names now read “<ticker ID>.csv”, for example, 1.csv.

In each CSV file, other than the file name itself, there was no indication of which ticker symbol or ID the price figures belonged to. Hence in this step, all the 941 ticker IDs were appended in a new column in the CSV files. This was to simplify the process of importing all the price figures into FDB-DS all at once without having it messed up while price figures were grouped according to each ticker ID.

5.5 Financial Discussion Boards Dataset (FDB-DS)

The design and creation of the FDB-DS depended upon all the identified artefacts (in Section 4.2) such as the ticker symbols, prices, comments, director deals and broker ratings from the FDBs. Figure 5.2 illustrates the novel design of the FDB-DS in the form of an Entity Relationship Diagram.

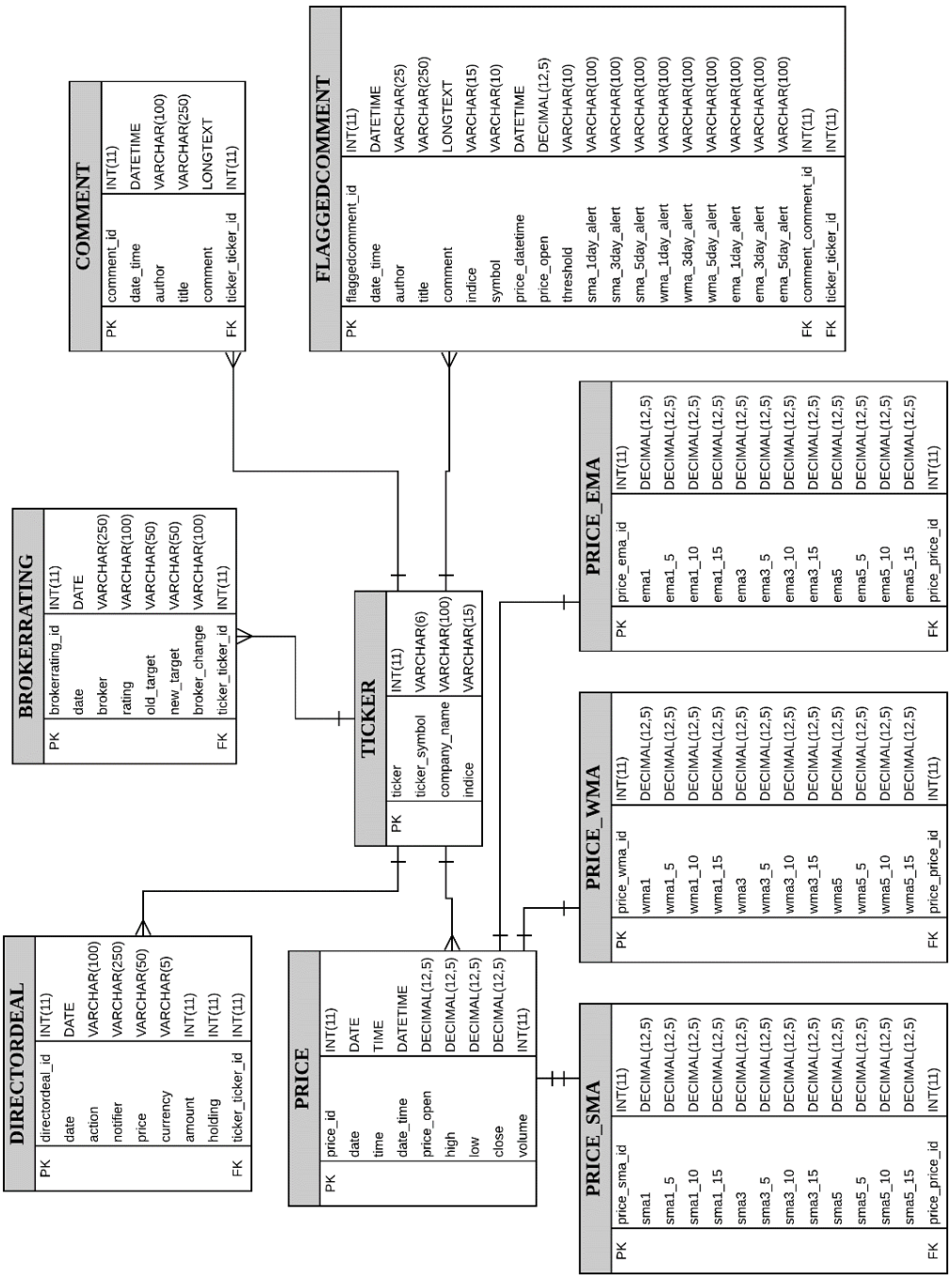

Figure 5.2 Entity Relationship Diagram (ERD)

Each database table is described as follows:

- **Ticker**

  Database table for storing all the 941 ticker symbols, company names and which index they belong to.

- **Comment**

  Database table for storing all the collected comments and the associated date, time, author, topic title, ticker ID and prices.

- **Flaggedcomment**

  Database table for storing the flagged comments when conducting experiments and analysis.

- **Price**

  Database table for storing all the prices and the associated date, time, moving average calculation and moving average threshold calculation outcomes.

- **Directordeal**

  Database table for storing the director deals artefact data and the associated date, time, amount and so on.

- **Brokerrating**

  Database table for storing all the broker and rating data and the associated date, time and target.

Each data type within the FDB-DS will now be described:

- **INT(size)**

  INT in a database means integer. In Figure 5.2, the primary key (i.e. `ticker_id`) in the `ticker` table is set with INT(12). “12” is the maximum number of digits being displayed in the database column.

- **DATE**

  DATE in a database means date only. The date data are stored in the FDB-DS using a “YYYY-MM-DD” format, e.g. 2014-11-30.

- **DATETIME**

  DATETIME in a database is used to store date and time in a same field. The date and time data stored in the FDB-DS uses a “YYYY-MM-DD HH:MI:SS” format, e.g. 2014-11-30 20:49:31.

- **VARCHAR(size)**

  VARCHAR in a database means variable character. A database table column with VARCHAR data type can hold numbers and letters. A VARCHAR field in MySQL database can hold up to a maximum of 65,535 characters.

- **LONGTEXT**

  LONGTEXT in a database is used to store long texts. Unlike VARCHAR, long text does not have a limit in terms of how many characters a field can store. However, LONGTEXT has a data size limit of 4GB. LONGTEXT is used for the data item such as `comment` since some FDB comments can be longer than the 65,535-character limit in VARCHAR.

- **DECIMAL(P,D)**

  DECIMAL in a database is used to store exact numeric values such as money data. "P" represents the precision that represents the number of digits and "D" represents the number of digits after the decimal point. In FDB-DS, "DECIMAL(12,5)" means a field can store up to 12 digits including the 5 digits after the decimal point, e.g. 1,234,567.89000.

5.6 Pump and Dump (P&D) Keyword Template

Table 5.3 demonstrates the list of P&D keywords, phrases and short sentences selected based on the methodology in Phase 4, described in Section 4.3.4. There is a total number of 117 keywords, phrases and short sentences in the P&D IE keyword template before stemming is performed. This list was validated by a human expert in the field through email communication.

Table 5.3 P&D IE Keyword Template

Keywords	Keywords
Pnd	Pump dump
Manipulate	Manipulating stock
Manipulating	Manipulating share
Manipulation	Chop stock
Manipulative	Penny stock

Misleading	False statement
Deceive	Misleading statement
Deceiving	Misleading positive statement
Deceptive	Once-in-a-lifetime
Hoax	Once in a lifetime
Scam	Pump the price
Falsify	Pump the share
Cheat	Pump this stock
Con	Inflated price
Spreading rumours	Hot stock
Artificially raising price	Huge volume spike
Artificially inflate price	Keep ramping
Pump	On hypes
Pumped	Buy now
Pumping	Good future
Ramp	Invested so heavily
Ramped	It will fly
Ramping	Buy more
Inflate	Rock bottom price
Inflated	Buy on dips
Inflating	Buy the dips
Tips	Best tips
Incline	Good tip
Elevate	Dump the price
Promote	Dump the share
Boost	Short stock
Get rich	Fall hard
Make a killing	Sell now
Make a fortune	Sell quickly
Make money	Declining stock
Make big bucks	Begin to decline
Make a pretty penny	Begin to deteriorate
Make a bundle	Volume is dead
Make a bomb	When I post you better believe me
Make a packet	Do not doubt me on this one
Hit it big	Anyone who sells to make a small profit is very short sighted
Gain profit	This is the chance
Become wealthy	Price will go up
Strike it rich	Buy as quickly as possible
Dump	Take what profit you can
Dumped	About to run, get in today
Dumping	Adding a few more don’t let the shorts win here
Short	Buying opportunity this AM
Shorted	Still looks good for a run
Shorting	This is the opportunity

Ditch	shall go up again after today
Drop	speculated return
Opt out	Now is the time
Abandon	Chance to make some real dollars
Back out	This stock might not have any trouble hitting over
Give up the ship	May be too late if we wait a bit longer
Pull out	Sell as quickly as possible
Let go	Get out while you can
Pump and dump	-

5.7 Forward Analyser

The aim of the forward analyser component is to filter and flag the potentially illegal P&D comments using the IE keyword template with the integration of share prices. The forward analyser contains a number of functions to operate through the entire forward comments flagging process. These functions include the initial flagging of potentially illegal comments, matching prices to flagged comments and price hike threshold labelling.

5.7.1 Comment Flagging

The forward analyser will initially perform a stemming process against the P&D IE keyword template and comments using the commonly used Porter's stemming algorithm (Porter, 1980). Then all the keywords, phrases and sentences from the P&D IE keyword template will be matched against the 507,970 comments which were pre-processed and transformed into structured data stored in the FDB-DS by the data transformer as described in Section 5.4. The results of the flagged potentially illegal comments will then be imported into a new table named `flaggedcomment` in the FDB-DS for the use of experiments and analysis in Chapter 5.

5.7.2 Price Matching

As mentioned in Section 5.3.2, though the per minute share prices for ticker symbols are available through an ADVFN monthly subscription, the full availability of all price

figures is not guaranteed. As a result of this, “0.00” is recorded when there is no price captured by ADVFN. These zero values must be handled prior to performing price matching in this section and in Section 5.8.1 for moving average calculation. When zero values occur in the prices of a ticker ID, pick the last known or next price figure to replace the zero values. This is to prevent any inaccuracy in the final outcomes of this research.

Once the flagged comments table has been populated, the forward analyser will proceed to the process of matching a price to each flagged comment. The method used at this stage attempts to set a “base price” for each comment, particularly the flagged comments, since thresholds will be applied in the next stage. At this stage, the analyser will read each comment’s date and time, then search for the price figure of each ticker ID that shares the same date and time of the comment of a specific ticker symbol. If a price figure at the same date and time is not found, the next available price figure will be appended to the flagged comment. Due to the extremely large `price` table – consisting of over 28 million rows – after all the testing, this process takes up to two weeks to completely search and match price figures for all the 49,858 rows of flagged comments.

5.7.3 Threshold Labelling

After having all the “base prices” set for each flagged comment according to the identical or nearest comment’s and price’s date and time, the analyser will attempt to label each flagged comment with thresholds. This process takes days to completely calculate all the prices within ±2 days against the “base price” and then sets the triggered threshold to each flagged comment. The threshold labelling rules are as follows:

- Flagged comments that have no price figure (due to incomplete price figures provided by ADVFN) and non-flagged comments will be labelled as “N” (Null).
- If any of the price figures from the prices within ±2 days exceeds a 5% price hike when calculated against the “base price” of a flagged comment, the comment will be labelled as “Y” (Yellow).

- If any of the price figures from the prices within ±2 days exceeds a 10% price hike when calculated against the “base price” of a flagged comment, the comment will be labelled as “A” (Amber).
- If any of the price figures from the prices within ±2 days exceeds a 15% price hike when calculated against the “base price” of a flagged comment, the comment will be labelled as “R” (Red).
- Flagged comments with a “base price” that does not exceed a 5% price hike will be labelled as “C”.

5.8 Backward Analyser

The aim of the backward analyser component is to further classify the flagged comments (resulting from forward analysis) using the abnormality in share prices which is calculated using moving average methods. The novel methodology used in the backward analyser component also attempts to resolve false positives for flagged comments.

Like the forward analyser, the backward analyser consists of a number of major functions. These functions run one after another in order to achieve the final outcome. This includes calculating the moving average price figures, determining and labelling whether there are alerts for the original price exceeding the moving average price, and finally backward alert labelling towards the existing flagged comments.

5.8.1 Moving Average Calculation

Moving average is one of the technical analysis methods that is often widely used by financial analysts to predict future price patterns and learn stocks’ behaviour and trends by studying historical price data (Dzikevičius and Šaranda, 2010). The most common moving averages used are the Simple Moving Average (SMA) and Exponential Moving Average (EMA). But, Weighted Moving Average (WMA) is also widely used depending on the analysis needs. Some researchers even use a moving average for predicting the rate of traffic congestion and road accidents (Raiyn and Toledo, 2014). However, it appears that there have been few to no attempts in past

research to use moving average as a financial crime tool to detect abnormal stock price movements.

The function of the backward analyser is to attempt to apply the SMA, WMA and EMA to the pre-processed prices. Firstly, it was decided that the time periods used for this experiment are 1 day, 3 days and 5 days. These time periods can be changed by the relevant authorities if needed. Next, calculate the SMA, WMA and EMA using their formulas. Section 5.8.1.1 explains the calculation of SMA; Section 5.8.1.2 explains the calculation of WMA and Section 5.8.1.3 explains the calculation of EMA.

5.8.1.1 Simple Moving Average (SMA) Calculation

Equation 1 shows the formula for SMA calculation.

Eq. 1. $$SMA = \frac{p_1+p_2+\cdots+p_n}{n}$$

where, *p* = price; *n* = time period.

To understand how SMA works, Table 5.4 illustrates a simple example. To achieve the calculation of a 5-minute SMA, the analyser will pick the prices for the first 5 minutes and sum up, then divide by 5.

For example, (16+17+17+10+17)/5 = 15.4

Table 5.4 SMA Calculation Example

Time	**8:00**	**8:01**	**8:02**	**8:03**	**8:04**	**8:05**	**8:06**	**8:07**	**8:08**
Price ($)	16	17	17	10	17	18	17	17	17
5-min SMA					15.4	15.8	15.8	15.8	17.2

Thus, in this example, the first SMA of the prices is 15.4. When the calculation moves on to the next row (i.e. next price) in the database `price` table, the original first price figure (i.e. $16 at the time of 8:00) is excluded and the new price (i.e. $18 at the time of 8:05) is taken into account in order to calculate the next SMA value (i.e. $15.8 next to the first SMA value of $15.4).

Since the chosen periods for this purpose are 1 day, 3 days and 5 days and for each day, the stock market movement is active for 8 hours (480 minutes), the SMA calculation for a period of 1 day starts with summing up all the prices for the first 480 minutes and dividing the sum by 480 to get the SMA. This calculation continues for all the prices for each ticker symbol.

The same steps are repeated for the periods of 3 days and 5 days. The results of these SMA calculations are recorded into new price tables in the FDB-DS.

5.8.1.2 Weighted Moving Average (WMA) Calculation

In a WMA calculation, a greater weight is given to the most recent price data because the more recent price is deemed to be more important. Equation 2 shows the formula for WMA calculation.

Eq. 2. $$WMA = \frac{(P*n + P(1)*n-1+\cdots P(n-1)*1)}{\left(n*\frac{(n+1)}{2}\right)}$$

where, *p* = price; *n* = time period.

To achieve the calculation for a, take, for example, a 5-minute WMA, then carry out the following steps:

i. Sum up all the weights i.e. 1+2+3+4+5 = 15
ii. Then formulate the weighting by using the weight of each price divided by the sum of weights. For example, 1/15, 2/15, 3/15 and so on.

iii. After this, multiply the price with the weighting to produce the weighted value. For example, ((17*(5/15))+(10*(4/15))+(17*(3/15))+(17*(2/15))+(16*(1/15)))

iv. Finally, the calculated WMA figure is the sum of all the weighted values. For example, 1.07+2.27+3.40+2.67+5.67=15.07

Table 5.5 shows an example of the WMA calculation in a clearer manner.

Table 5.5 WMA Calculation Example

Time	8:01	8:02	8:03	8:04	8:05
Price ($)	16	17	17	10	17
Weight	1	2	3	4	5
Weighting	1/15	2/15	3/15	4/15	5/15
Weighted value	1.07	2.27	3.40	2.67	5.67
5-min WMA					15.07

5.8.1.3 Exponential Moving Average (EMA) Calculation

Like WMA, Exponential Moving Average (EMA) also add weights to the formula, so that the latest price data have more influence in the moving average charting. EMA is a moving average technique that is similar to SMA. The only difference is that EMA has weighting added and SMA does not. The weights in EMA are slightly different from the weights added in the WMA formula. The weights in WMA are consistently being decreased, but the weights in EMA are exponential, meaning they decrease at a greater rate instead of consistently like for WMA. In this case, EMA reacts to the latest prices more quickly (Ross, 2017).

To achieve the calculation of a 5-minute EMA, there is a need to first convert time period (n) to “K” (i.e. smoothing constant), then carry out the EMA calculations.

The formula to calculate EMA is as below:

Eq. 3. $$EMA_{[today]} = (P_{[today]} \, x \, K) + (EMA_{[yesterday]} \, x \, (1 - K))$$

where, *p* = price; *K* = 2 / (n + 1); *n* = time period.

An example of the calculation is shown below, also in Table 5.6:

(17 * 2 / (5 + 1)) + [16.3 * (1 – 2 / (5 + 1))]

= (17 * 0.3333) + [16.3 * (1 – 0.3333)]

= 5.6661 + 10.86721 = 16.5

Table 5.6 EMA Calculation Example

Time	8:01	8:02	8:03	8:04	8:05	8:06	8:07	8:08	8:09
Price ($)	16	17	17	10	17	18	17	17	17
33.3% EMA			16.5	14.4	15.2	16.2	16.4	16.6	16.8

5.8.2 Alert Labelling

After obtaining all the SMA, WMA and EMA values in the previous step, three thresholds are applied to all the moving average values (which act as the moving average base value). The thresholds are 5%, 10% and 15% above the moving average values.

For example, as shown in Table 5.7, if an SMA value is $15.4, all the three threshold values are as below:

Table 5.7 Moving Average 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

The purpose of these calculated and recorded thresholds is meant for the next section where the threshold alerts are matched backwards to the potentially illegal comments.

5.8.3 Alert Matching

In the FDB-DS, there will be a database table that records the flagged potentially illegal comments by performing the comments flagging process described in the next chapter. To recap on how it reaches this stage, the comments will be flagged initially through the process of keyword matching using the P&D IE keyword template. Then the flagged comments will be appended with the exact or nearest price value at the time each comment was posted. This price then acts as a "base price" for the forward analyser to count it against all the prices of ±2 days of a ticker symbol, in order to determine and label them with price hike thresholds.

After all the steps (Section 5.7.1, Section 5.7.2 and Section 5.7.3) in the forward analyser component and moving average calculations (as stated in Section 5.8.1.1, Section 5.8.1.2 and Section 5.8.1.3) are performed, the backward analyser will then investigate backwards by matching the price thresholds to the flagged comments based on the same or nearest date shared between flagged comments and the prices.

5.9 Graphical User Interface (GUI)

The GUI for FDBM is simple as it is mainly used for managing button click events which execute backend functions, retrieving and displaying data from the FDB-DS. Section 5.9.1 describes and illustrates the GUI for the data crawler and Section 5.9.2 demonstrates the GUI for the forward and backward analysers.

5.9.1 Data Crawler GUI

The data crawler component is designed to be fully automated when crawling for data. The other executable functions are also meant to operate in the background. However, the crawler GUI is designed for just in case there is a need to manually

initiate a download process for specific data, for example, fetching director deals data.

In the data crawler GUI, the buttons like “Fetch RNS” and “Fetch Financial Diary” contain executable functions. Though these data are not collected and not used in the current research due to challenges like time and resources, it can be useful for future research (which is why these functions are not removed at the time of writing this).

Figure 5.3 and Figure 5.4 illustrate the data crawler GUI in two parts. Each button and each data display table have been labelled as follows:

1. **Fetch Tickers**

   To download ticker symbols manually.
2. **Load Tickers**

   To load the downloaded ticker symbols from the FDB-DS.
3. **Fetch Broker Ratings**

   To download broker rating data manually.
4. **Load Broker Ratings**

   To load the downloaded broker rating data from the FDB-DS.
5. **Fetch Recent Trades**

   To download recent share trades data manually.
6. **Load Recent Trades**

   To load the downloaded recent share trades data from the FDB-DS.
7. **Fetch Director Deals**

   To download director deals data manually.
8. **Load Director Deals**

   To load the downloaded director deals data from the FDB-DS.
9. **Fetch Financial Diary**

   To download financial diary data manually.
10. **Load Financial Diary**

    To load the downloaded financial diary data from the FDB-DS.
11. **Fetch RNS**

    to download RNS (regulatory news) manually.

12. Load RNS

To load the downloaded RNS from the FDB-DS.

13. Ticker Symbols Data Display

The loaded ticker symbols will be displayed in this area.

14. Recent Trades Data Display

The loaded recent trades data will be displayed in this area.

15. Financial Diary Data Display

The loaded financial diary data will be displayed in this area.

16. Broker Ratings Data Display

The loaded broker rating data will be displayed in this area.

17. Director Deals Data Display

The loaded director deals data will be displayed in this area.

18. Regulatory News Data Display

The loaded RNS will be displayed in this area.

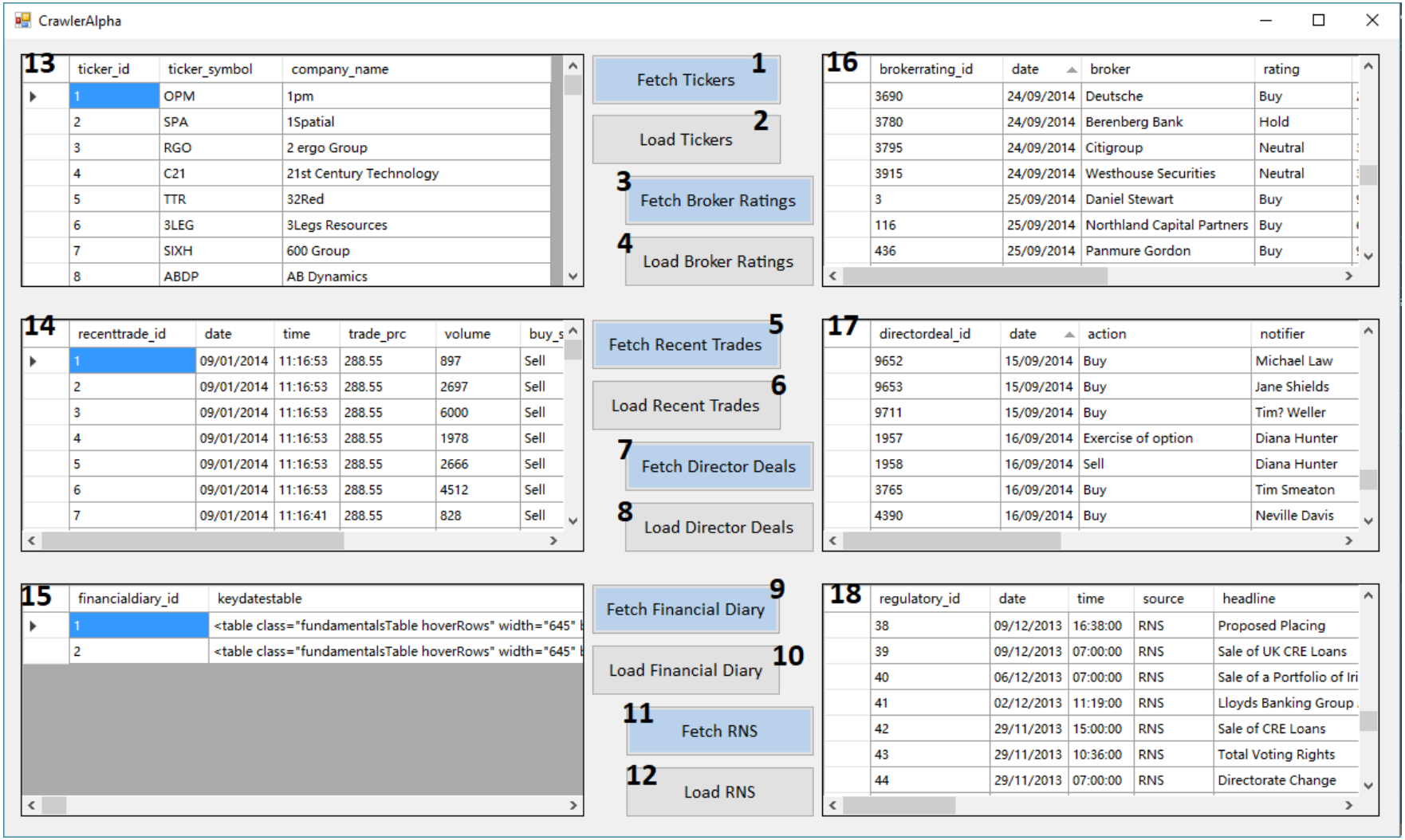

Figure 5.3 Data Crawler GUI – Part 1

19. Fetch Fundamentals

To download fundamentals data manually.

20. Load Fundamentals

To load the downloaded fundamentals data from the FDB-DS.

21. Fetch Share News

To download share news data manually.

22. Load Share News

To load the downloaded share news data from the FDB-DS.

23. Fetch Share Prices

To download share prices data manually.

24. Load Share Prices

To load the downloaded share prices data from the FDB-DS.

25. Fetch Comments

To download comments data manually.

26. Load Comments

To load the downloaded comments data from the FDB-DS.

27. Fundamentals Data Display

The loaded fundamentals will be displayed in this area.

28. Share News Data Display

The loaded share news data will be displayed in this area.

29. Share Prices Data Display

The loaded share prices data will be displayed in this area.

30. Comments Data Display

The loaded comments data will be displayed in this area.

31. Ticker Symbols Data Display

The loaded ticker symbols data will be displayed in this area.

32. Load Tickers

To load the downloaded ticker symbols from the FDB-DS.

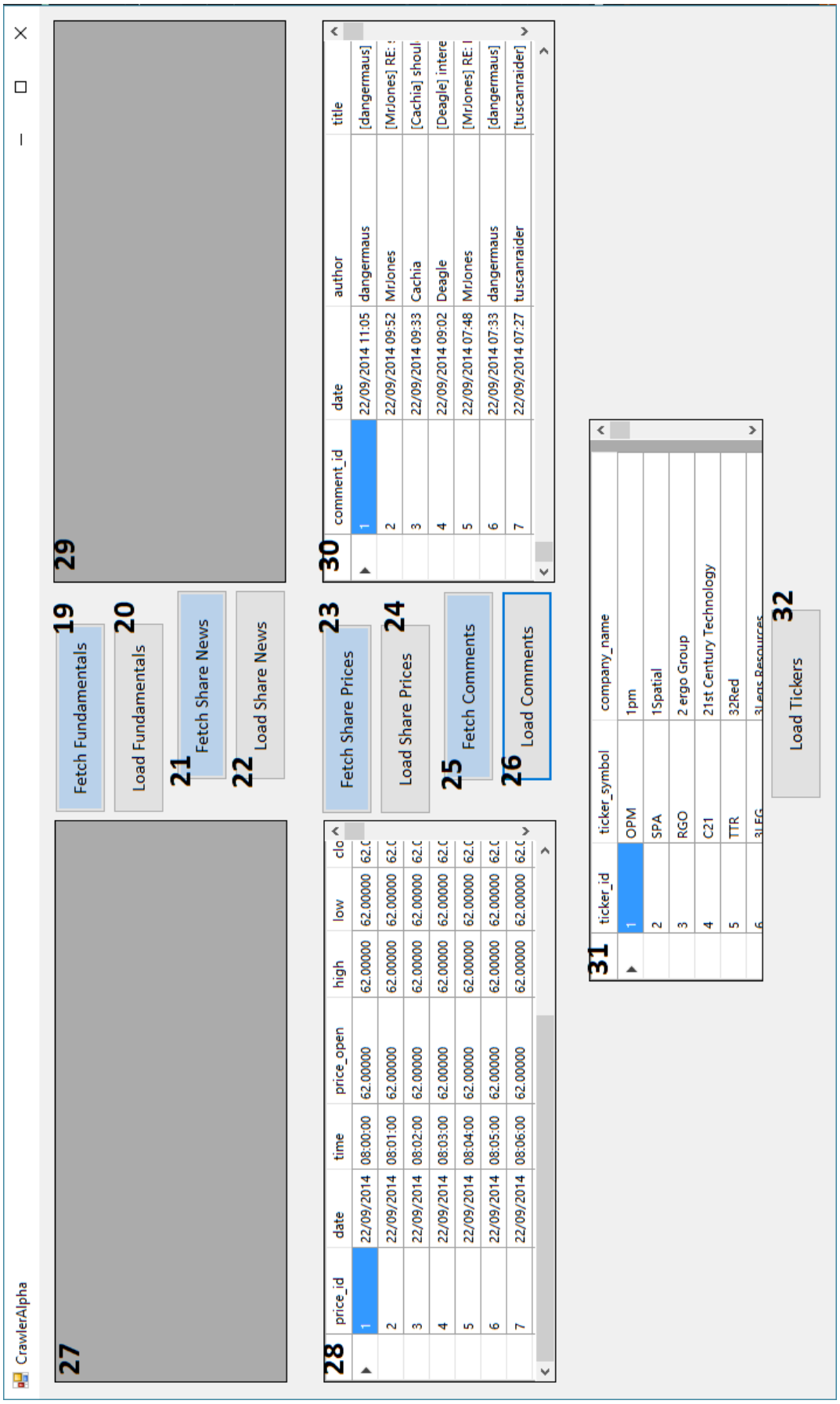

Figure 5.4 Data Crawler GUI – Part 2

5.9.2 Forward Analyser GUI

The forward analyser is also comprised of many background functions which hide behind one-click buttons. As such, the frontend GUI is designed to focus more on retrieving data and displaying the data from the FDB-DS. Figure 5.5 shows the GUI of the forward analyser.

Figure 5.5 Forward Analyser GUI

The functions, which mostly work in sequence, on the forward analyser GUI are described below:

1. **Keyword Set Dropdown Menu**

   This allows the user to choose which financial crime IE keyword template to use, provided they have been predefined.

2. **Flagged Comment (Data View)**

   After a keyword template has been selected and the Search button has been hit, the comments flagging (as described in Section 5.7.1), price matching (as described in Section 5.7.2) and the price hike thresholds labelling (as described in Section 5.7.3) will start and the results of flagged comments will appear in this as the data view.

3. **Flagged Comment (Text View)**

   The results appearing in the Data View will also appear here as text format.

4. **Select Stock Index**

   This dropdown list will be populated according to the flagged comments. For example, if any of the comments belonging to the LLOY ticker symbol are flagged, LLOY will appear in this dropdown list for the user to select. Selecting LLOY will update both the Flagged Comments (Data View) and Flagged Comment (Text View) to show only the flagged comments which belong to LLOY.

5. **Select Date**

   The dropdown list here allows the user to choose the date of the flagged comments belonging to LLOY.

6. **Total Comments Count**

   This field will change dynamically according to the count of flagged comments appearing in the Flagged Comment (Data View).

7. **Prices (±2 Days Intervals)**

   Once the user has selected the date from the Select Date dropdown list corresponding to the selected Stock, ±2 days' worth of prices based on the date will appear.

8. **Thresholds**

   By clicking on any of these buttons the Flagged Comments (Data View) and Flagged Comments (Text View) will be filtered. This means if the user wants to have a look at all the comments that are being labelled with a price hike threshold of 5% (Y), the user will click the “5%” button. The same action goes for 10% and 15%. Finally, if the user wants to show all the flagged comments regardless of the price hike thresholds again, the “Show All” button should be clicked.

9. **Update Data View**

   This button will allow the user to click in order to update the Flagged Comments (Data View) after adding notes to one column that are editable in the data view.

10. **Import Into Database**

    The updated Flagged Comments (Data View) can be imported into the database by clicking on this button.

11. **Load Flagged Comments**

    This allows the user to reload the Flagged Comments (Data View) alongside Flagged Comments (Text View). This can be used when there is no forward analysis being done. It merely loads the flagged comments from the `flaggedcomment` table from the Financial Discussion Boards Dataset (FDB-DS).

12. **Random**

    By clicking on this button, the flagged comments from `flaggedcomment` will be randomly chosen, and by default the limit is set to 1,000 rows of data. This might be handy if the user just wants to randomly pick some flagged comments for manual analysis by manually reading them.

13. **Export Text View**

    This function allows the user to export the Flagged Comments (Text View) into a text file. This can be useful if the user wants to print the flagged comments based on the criteria set using certain functions introduced above.

5.9.3 Backward Analyser GUI

The backward analyser allows the user to execute moving average calculations (as described in Section 5.8.1), alert labelling (as described in Section 5.8.2) as well as alert matching (as described in Section 5.8.3) backwards to the flagged comments. Similar to the forward analyser GUI, the functions are in the background. Hence, the backward analyser GUI provides the user with options to retrieve and view data. Figure 5.6 represents the GUI of the backward analyser.

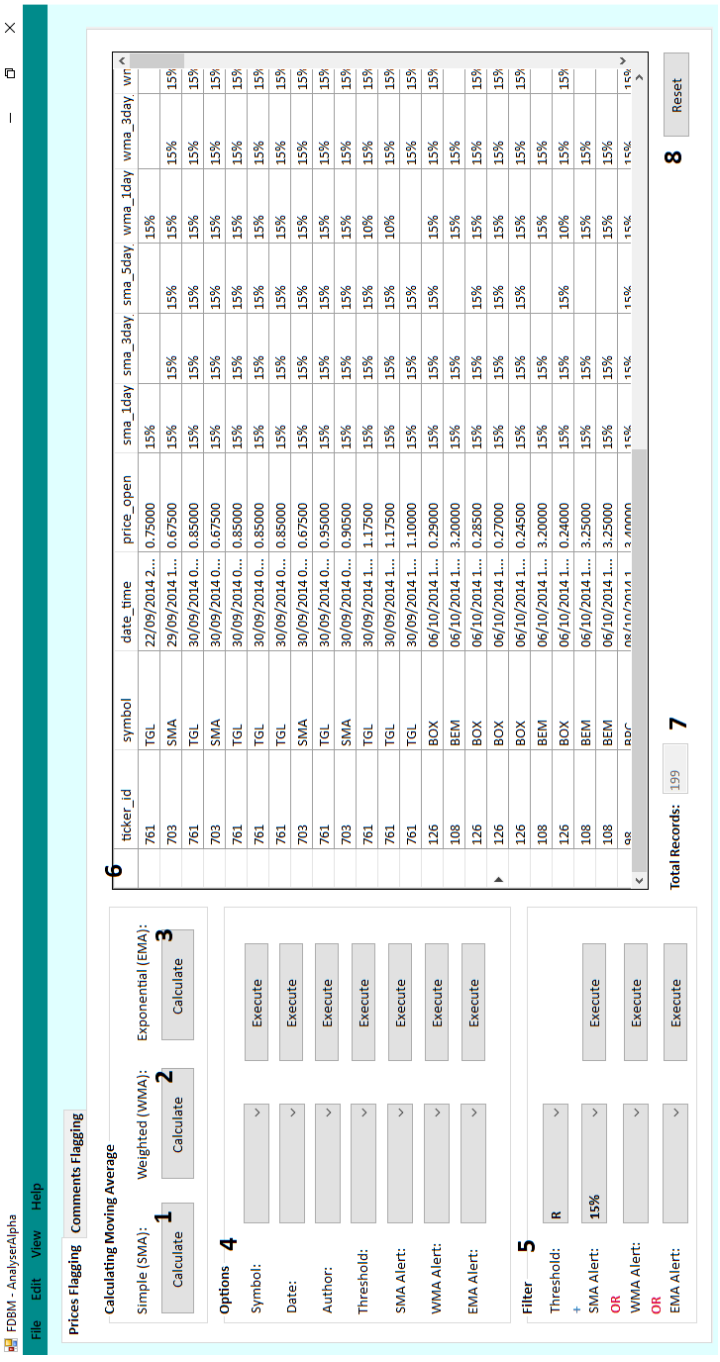

Figure 5.6 Backward Analyser GUI

The functions on the backward analyser GUI are described below:

1. **Calculate (for SMA)**

   By clicking this button, all the functions of the backward analyser, i.e. moving average calculation, alert labelling and alert matching will be executed for SMA.

2. **Calculate (for WMA)**

   By clicking this button, all the functions of the backward analyser, i.e. moving average calculation, alert labelling and alert matching will be executed for WMA.

3. **Calculate (for EMA)**

   By clicking this button, all the functions of the backward analyser, i.e. moving average calculation, alert labelling and alert matching will be executed for EMA.

4. **Options**

   This section allows the user to filter the data in the Data View. For example, the user can choose a comment author's FDB username from the Author dropdown list and view all the flagged comments posted by this particular author.

5. **Filter**

   This section allows the user to combine the threshold and moving average alerts. Once executed, the flagged comments will be listed in the Data View according to the selected filters. For example, if the user chooses the "R" threshold and SMA Alert of "15%", in the Data View, the user will be shown a list of flagged comments that match both filter criteria.

6. **Data View**

   An area for displaying the filtered flagged comments data as a result of the Options and Filter functions.

7. **Total Records**

   The field here will change dynamically depending on how many rows of data appear in the Data View.

8. **Reset**

   The reset button allows the user to reset all the dropdown lists and the Data View.

5.10 Chapter Summary

Section 5.2 described and presented the novel architecture of the FDBM prototype system which implemented the methodologies described in Section 4.3. Section 5.3 to Section 5.8 expanded on the architecture components in relation to the methodologies. Lastly, Section 5.9 displayed the GUI for the data crawler component, forward analyser component and the backward analyser component.

The key contributions of the FDBM prototype system introduced in this chapter are listed as follow:

- FDBM is capable of automatically crawling data in different formats and different time intervals from the selected FDBs.
- FDBM can pre-process and transform the unstructured FDB data collected from FDBs into structured data for importing into the FDB-DS for the use of forward analysis and backward analysis.
- FDBM is capable of matching the predefined P&D IE keyword template against the comments in order to detect the potentially illegal FDB comments.
- In order to categorise the potentially illegal FDB comments into different price hike levels at a later stage, it is essential to know the price at the time of each comment being posted. However, during the collection of FDB comments, there were no prices attached to each comment. Hence, FDBM is built to be capable of searching for the price figure at the time a comment is posted and appending it to each comment.
- FDBM is capable of performing the novel backward analysis by executing the calculation of all the prices within ±2 days against the “base price” of each flagged comment in order to categorise and label the comments according to the degree of price hike on three levels of thresholds (i.e. 5%, 10% and 15%).

- As part of the novel backward analysis, FDBM can also calculate and apply the moving averages against all the per minute share prices for all the ticker symbols, append threshold alerts and finally match backward to further classify the potentially illegal FDB comments in an attempt to resolve false positives.

The next chapter will present four experiments (two experiments under forward analysis and two experiments under backward analysis). The purpose of these experiments is to test and validate the novel forward analysis and backward analysis methodologies introduced in Chapter 4 and the architecture proposed in this chapter, which is in relation to the research question outlined in Chapter 1.

Chapter 6. Forward and Backward Analysis of Potentially Illegal Comments

6.1 Introduction

In Chapter 5, the architecture for a Financial Discussion Boards Miner (FDBM) prototype system has been described and FDBM's crawler component has been used to extract comments from London South East (LSE-FDB) and Interactive Investors (III) Financial Discussion Boards (FDBs) and per minute share prices from ADVFN. The chapter also described how data were collected using FDBM. The data were collected over a 12-week period using a data crawler component, pre-processed through a data transformer component and stored in the database, namely the FDB dataset (FDB-DS).

In order to answer the research question outlined in Chapter 1, a series of experiments were conducted using both novel forward and backward analysis methodologies. This chapter looks at both forward analysis and back analysis of the data stored in the FDB-DS. The aim of the forward analysis is to flag and filter the potentially illegal Pump and Dump (P&D) comments using the predefined P&D Information Extraction (IE) keyword template and the share prices. And, the aim of the backward analysis is to detect abnormal stock price movements and perform backward analysis by matching abnormal stock prices with the flagged comments to further classify flagged comments with the intention of reducing false positives for flagged comments.

This chapter consists of four experiments. The first two experiments are under forward analysis (Sections 6.2 and 6.3) and the next two experiments are under backward analysis (Sections 6.4 and 6.5). For forward analysis, the first experiment takes only the comments into account when flagging potentially illegal comments using the P&D IE keyword template. The second experiment takes share prices into account. The share prices at the time or around the time of these flagged comments (results produced in the first experiment) were posted are appended. This allows the calculation of price hike thresholds and allows these thresholds to be labelled for the

flagged comments. Doing this can achieve the classification of flagged comments into different risk levels, such as high risk, medium risk and low risk. It then allows the relevant authorities or forum moderators to first focus on the highest risk flagged comments.

For backward analysis, the third experiment in this chapter applies Moving Average (MA) techniques to the share prices, to determine whether there are price abnormalities. Lastly, the fourth experiment attempts to apply these price abnormalities backwards to the flagged comments produced in the forward analysis experiments, in order to further classify the flagged comments in an attempt to reduce false positives for flagged comments. Abnormal price movement happened when the price of stock varied from the market average.

6.2 Experiment 1: Forward Analysis of FDB Comments

The aim of this experiment is to test whether the P&D IE keyword template can be used to flag potentially illegal P&D comments on share price based FDBs.

This experiment will test the following hypothesis:

H_{0a}: Pump and Dump activity from FDBs cannot be filtered using only information (i.e. keywords/phrases) extracted from collected FDB comments.

H_{1a}: Pump and Dump activity from FDBs can be filtered using only information (i.e. keywords/phrases) extracted from collected FDB comments.

6.2.1 Methodology

The following describes the steps in the methodology implemented by FDBM to test the hypothesis:

1. Retrieve the predefined P&D IE keyword template from the local storage of the FDBM prototype system.

2. Iterate through each of the keywords, phrases and sentences in the IE keyword template (Section 5.6) against each of the comments stored in the dataset (FDB-DS).
3. Flag the comments that matched any of the keywords, phrases or sentences.
4. Import these flagged comments into a new database table, named `flaggedcomment`, in the FDB-DS, for subsequent experiments.
5. List the results in two views, i.e. data view and text view (for export into a text file if required by users of FDBM prototype system).

6.2.2 Results and Discussions

Out of the 507,970 comments, 49,858 comments are flagged as potentially illegal P&D activities on the FDBs. As shown in Figure 6.1 below, the flagged comments took up 9.82% of the total comments.

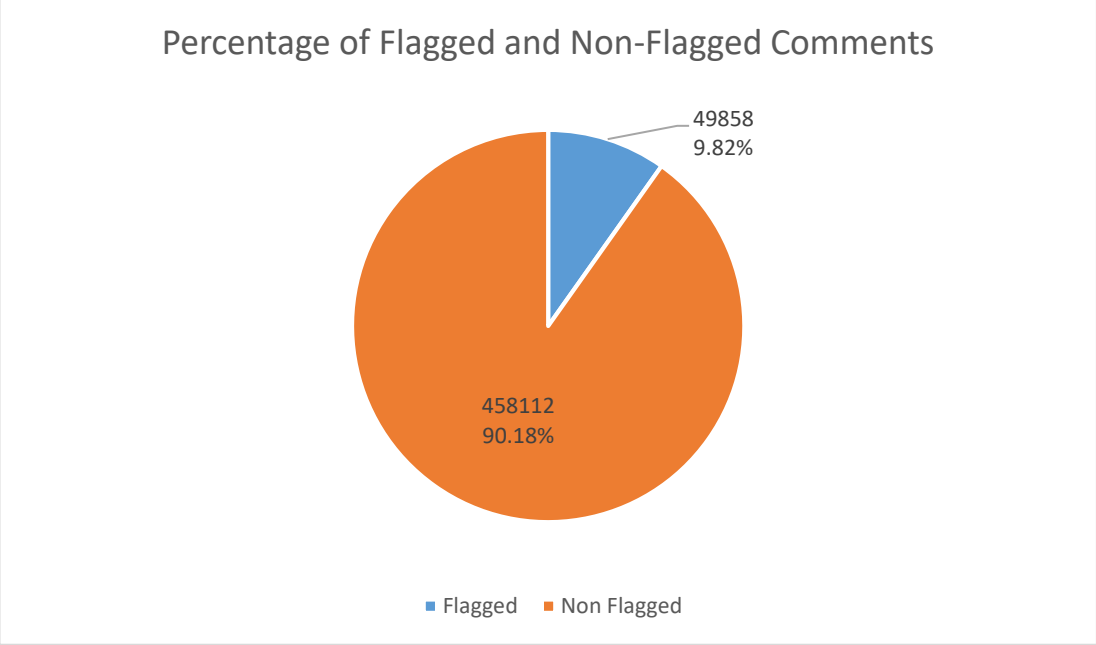

Figure 6.1 Percentage of Flagged and Non-Flagged Comments

As shown in Figure 6.2 below, of these 49,858 flagged comments, 46,302 (93%) of the comments belong to the FTSE AIM All-Share index and 3,556 (7%) of the comments belong to the FTSE-100.

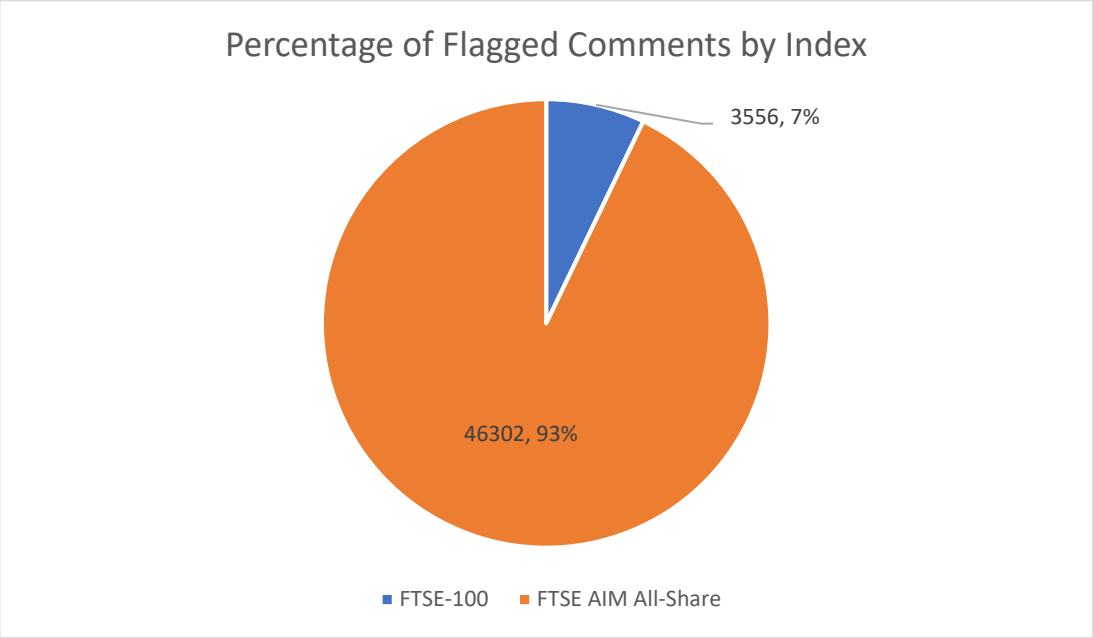

Figure 6.2 Percentage of Flagged Comment by Index

The following are a few examples of the flagged comments:

Table 6.1 Examples of Flagged Comments

FC1	**Ticker ID**: 530 **Ticker Symbol**: ODX **Author**: Hufc1908 **Date/Time**: 2014-09-22 04:52:56 **Comment**: I know Doc & he is very level headed guy UJO & MXO both good small AIM tiddlers nice to see Adam doing a bit of ramping :) Here is Docs tweet from last night & please don't say anyone here holds SLE?? The chairman is a disgrace just like someone else i could mention called Jim who pulled the same trick as COE of Sefton ResI've been reporting what a shister this man is for yonks #SLE #FANNING @TomWinnifrith @BrokermanDaniel @ABMckinley pic.twitter.com/d8kJJN5S8B
FC2	**Ticker ID**: 178 **Ticker Symbol**: COMS **Author**: englishflowers **Date/Time**: 2014-09-22 07:49:00

	Comment: Auction has just finished now 4.83p Seems that there will be some good buying now at this price. GLA.
FC3	**Ticker ID**: 616 **Ticker Symbol**: QXT **Author**: canapa123 **Date/Time**: 2014-09-22 09:06:29 **Comment**: Looks like some manipulation going on this morning with a good dash of trolling. I hope the FCA are watching. Hold on to your shares & you will be well rewarded imo.
FC4	**Ticker ID**: 616 **Ticker Symbol**: QXT **Author**: PJM1 **Date/Time**: 2014-12-19 09:43:06 **Comment**: listen.....fill yer boots it will be 37 again soon trust me

Some flagged comments are not potentially illegal themselves. However, these comments can be a great indication of P&D activities going on for a specific ticker symbol in a specific timeframe. In a real-world scenario, instead of moderating comment by comment, the FDBM prototype system can potentially help relevant authorities and human moderators to detect which listed company is being abused for conducting such financial crime. For example, FC1 in Table 6.1 shows that the comment author "*Hufc1908*" for the Ticker ID of 530 (Company name: Omega Diagnostics Group; Symbol: ODX) claimed that "…*nice to see Adam doing a bit of ramping…*". FC3 shows the comment author "*canapa123*" commented "*Looks like some manipulation going on this morning with a good dash of trolling. I hope the FCA are watching...*" for the Ticker ID of 616 (Company name: Quixant; Symbol: QXT). This does not necessarily indicate the comment itself is an illegal comment, but it can raise an alert about potentially illegal activities going on. On another hand, FC2 and FC4 show direct potentially illegal P&D comments. In FC2, the comment author "*englishflowers*" was potentially trying to create a good impression for the particular stock Ticker ID 178 (Company name: Coms; Symbol: COMS) by saying "*Seems that there will be some good buying now at this price*". Genuine sentiment or not, such a comment is an indication of claims without solid proof. In FC4, the comment author "*PJM1*" was trying to convince other investors to "*trust me*" that the share price of Ticker ID 616 will be "*37*" again. "*37*" in this case means 37 pence sterling (i.e. GBX being a symbol for the penny used in the London Stock Exchange (LSE)). Therefore, the results of this experiment support hypothesis H_{1a}.

Potentially illegal P&D activity on FDBs can be filtered using only information (i.e. keywords, phrases and sentences) extracted from collected FDB comments. However, in the real world, the number of flagged comments that must be reviewed is significantly high. On average, for just half of the listed companies on the LSE, a relevant authority or a forum moderator needs to read through 593 flagged comments each day, which still requires massive effort and is time consuming. Hence, Experiment 2 in the next section is performed, this time with the involvement of share prices instead of just comments.

6.3 Experiment 2: Forward Analysis of FDB Comments and Prices

The aim of this experiment is to examine whether the share prices can be taken into account in order to filter the flagged comments according to the price movements and the indices they belong to.

This experiment tests the following hypothesis:

H_{0b}: Pump and Dump activity from FDBs cannot be filtered using the IE keyword template and their correlation with price movements and the index they belong to.

H_{1b}: Pump and Dump activity from FDBs can be filtered using the IE keyword template and their correlation with price movements and the index they belong to.

According to an event study conducted by Bouraoui (2015), when a P&D event happens, the share price on the event day hikes and gradually decreases over the next two days. As discussed in Chapter 2, Leung and Ton (2015) who conducted an event study on the Australian Stock Exchange (ASX), have also discovered that the average abnormal returns for the stock prices have already increased significantly two days before the P&D event day. Sabherwal et al. (2011) also found the same pattern in their P&D research related to the NASDAQ stock exchange in the US, where there was a two-day pump followed by a two-day dump in the stock prices. Thus, it

was decided that all the ±2 days (a total of 5 days) of the per minute share prices should be taken into account when calculating the price hike thresholds for appending to the flagged comments in Experiment 2.

The outcome of price hike threshold calculations is categorised into the following five categories:

- R – R (red) represents the flagged comments that have exceeded the price hike threshold of 15% and above.
- A – A (amber) represents the flagged comments that have exceeded the price hike threshold of 10% but are below 15%, 10% to 14.99%.
- Y – Y (yellow) represents the flagged comments that have exceeded the price hike threshold of 5% but are below 10%, i.e. 5% to 9.99%.
- C – C represents the flagged comments that do not exceed the price hike threshold of 5%, i.e. 0% to 4.99%.
- Null – Null represents the flagged comments that were not able to be calculated for the price hike thresholds due to missing "base price". The main reason for this is that, when collecting per minute share prices from ADVFN, the prices were already missing.

6.3.1 Methodology

The following describes the steps taken to test the hypothesis:

1. Retrieve the flagged comments result from `flaggedcomment` database table in the FDB-DS.
2. Based on the same or nearest date and time, match and record the share prices to each flagged comment.
3. Once all the flagged comments and prices are matched, take all the ±2 days' worth of per minute prices, calculate each price figure against the "base price" (i.e. the price of each flagged comment).
4. Determine whether there are any price figures (from the ±2 days) that triggered the thresholds of 5% (Y) increment, 10% (A) increment and 15% (R) increment.

5. Depending on the threshold trigger, label each flagged comment with “Y”, “A” or “R”. Label the flagged comments as “C” when the prices do not exceed 5%.

6.3.2 Results and Discussions

As illustrated in Figure 6.3 and Table 6.2, of all the 49,858 flagged comments, 37,895 (76.01%) flagged comments did not trigger Y, A and R thresholds, i.e. as explained, threshold “C” represents flagged comments that do not even hit threshold “Y”. There are 5,197 (10.42%) flagged comments labelled as Y (±2 days’ prices exceeded 5% threshold), 2,555 (5.12%) flagged comments labelled as A (±2 days’ prices exceeded 10% threshold) and 3,613 (7.25%) flagged comments labelled as R (±2 days’ prices exceeded 15% threshold). Null represents the flagged comments that failed to be appended with prices, mainly due to missing share price data in ADVFN. Thus no price hike threshold can be counted.

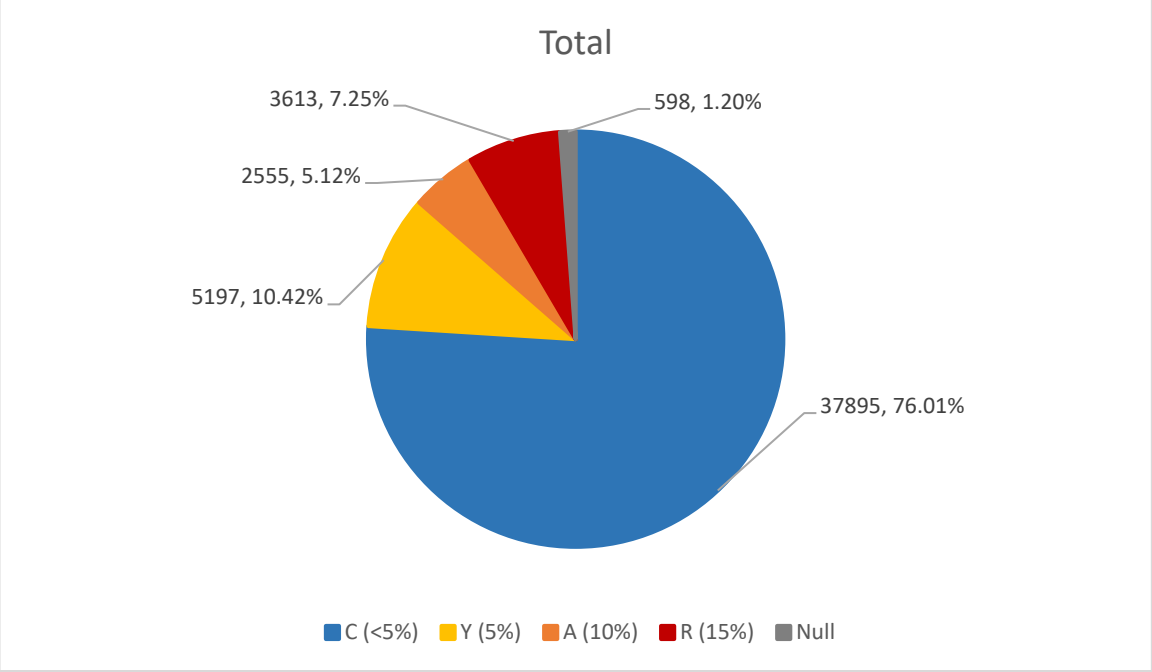

Figure 6.3 Total Number and Percentage for Each Threshold

Table 6.2 Total Number and Percentage for Each 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	1.20%
Grand Total	49,858	100%

After looking at the overall results, here comes the interesting part in relation to the indices on the LSE (LSE, 2017). As mentioned earlier in Chapter 1, the reasons for choosing the FTSE-100 and FTSE AIM All-Share are to observe which index is more prone to abuse by P&D fraudsters. It appears that, as revealed in Table 6.3, only 25 (0.7%) flagged comments in relation to the FTSE-100 are labelled as "R" (15% price hike) and only two (0.06%) of the flagged comments are being labelled as "A" (10% price hike). Even the flagged comments labelled as "Y" has only 55 (1.55%) of them and the remaining of the 3,337 (93.84%) flagged comments are categorised as "C", which represents non-risky flagged comments.

Table 6.3 Number and Percentage of Flagged Comments Based on Indices

	FTSE-100		FTSE AIM All-Share	
C (<5%)	3,337	93.84%	34,558	74.64%
Y (5%)	55	1.55%	5,142	11.10%
A (10%)	2	0.06%	2,553	5.51%
R (15%)	25	0.70%	3,588	7.75%
Null	137	3.85%	461	1.00%
Total	3,556	100.00%	46,302	100.00%
Grand Total	49,858 flagged comments			

On the other hand, the flagged comments belonging to the FTSE AIM All-Share index have been labelled with almost 20% fewer items in the non-risky threshold "C" than those for the FTSE-100. This means more FTSE AIM All-Share flagged comments are

worth investigating by relevant authorities, under the thresholds of “Y”, “A” and “R” with a total of 34,558 (74.64%), 5,142 (11.1%) and 3,588 (7.75%) respectively. Overall, 461 of the flagged comments are labelled with Null, which accounts for only 1%.

The results produced in this experiment appear to support hypothesis H_{1b}. By applying the price hike thresholds, it is possible to filter potentially illegal FDB comments for a more in-depth analysis.

6.4 Statistical Test and Summary of Forward Analysis Experiments

The results of both Experiment 1 and Experiment 2 appear to support the hypotheses H_{1a} and H_{1b} respectively. It was found that, in Experiment 1, comments can be flagged by using just the P&D IE keyword template. However, due to the enormous amount of flagged comments, share prices needed to be taken into account in order to filter the flagged comments in relation to the price movements. The data produced through the experiments were validated through statistical tests.

The chi-square test was used for conducting a statistical test for the outcome of the forward analysis experiments. A chi-square test tests the relationship between two categorical variables (such as potentially illegal comments and non-potentially illegal comments) and ordinal data (such as 5%, 10% and 15%). The results of the chi-square test are shown in the following Table 6.4.

Table 6.4 Chi-Square Test

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	915.862[a]	3	.000
Likelihood Ratio	1355.096	3	.000
Linear-by-Linear Association	731.663	1	.000
N of Valid Cases	49858		

It shows that the p-value of the statistic test is significant i.e. chi-square=915.862, df =3, p-value<0.01. The P-value is observed in the experiments to determine whether there is a strong evidence against the null hypothesis. In these experiments, a small p-value (i.e. p-value<0.01) represents strong evidence for rejecting the null hypothesis and accepting the alternate hypothesis. This means that the number of flagged comments is affected by the prices and indices.

Table 6.5 Correlations

		Index	Threshold
Index	Pearson Correlation	1	.121**
	Sig. (2-tailed)		.000
	N	49858	49858
Threshold	Pearson Correlation	.121**	1
	Sig. (2-tailed)	.000	
	N	49858	49858

**. Correlation is significant at the 0.01 level (2-tailed).

The value of the Pearson correlation coefficient as shown in Table 6.5 is 0.121, which is positive and statistically significant. This also proves that the indices and the flagged comments are positively related. Table 6.3 in the previous section shows that in the FTSE-100, a total of 97.69% (i.e. 93.84% from “C” and 3.85% from “Null”) messages correspond to less than the 5% threshold. But, in the FTSE AIM All-Share there are about 20% fewer; it contains only 75.64% (i.e. 94.64% from “C” and 1% from “Null”) messages that correspond to less than the 5% threshold. This means that potentially relevant authorities can even concentrate their investigations on the FTSE AIM All-Share index alone.

The purpose of the backward analysis experiments in the next sections is to determine whether the moving average techniques could assist the relevant authorities better in terms of confirming whether a flagged comment is indeed an indication of P&D crime on FDBs.

6.5 Experiment 3: Backward Analysis of Prices

The aim of this experiment is to detect abnormalities in the stock price movements using moving average techniques without relating them to the flagged comments.

This experiment tests the following hypothesis:

H_{0c}: It is not possible to detect abnormal stock price movements by using moving average techniques.

H_{1c}: It is possible to detect abnormal stock price movements by using moving average techniques.

As discussed in Section 5.8.1, moving average is one of the statistical methods that is often widely used by financial analysts to perform financial related studies (Dzikevičius and Šaranda, 2010). The most common moving average used is the Simple Moving Average (SMA), followed by the Exponential Moving Average (EMA) as it is similar to the SMA but with "weights" so that the more recent prices are taken into account with a higher importance. Weighted Moving Average (WMA) also uses the concept of "weights". The decrement of weights in WMA is consistent, whereas the decrement of weights in EMA is exponential, meaning that they do not decrease in a consistent manner but faster. SMA, WMA and EMA are all widely used depending on the types of analysis being performed. Some researchers even use a moving average for predicting the rate of traffic congestion and road accidents (Raiyn and Toledo, 2014).

It appears that there have been no attempts in previous research known to the author to use the moving average as part of a financial surveillance tool to detect abnormal stock price movements in relation to the detection of potentially illegal FDB comments. This experiment attempts to prove the hypothesis H_{1c} by applying the SMA, WMA and EMA, as well as the price hike thresholds (i.e. 5%, 10% and 15%) to the prices in order to identify meaningful discoveries such as whether there are abnormalities in the price movements without the intervention of flagged comments.

A total of 28,980,465 per minute share prices that belong to all the 941 ticker symbols from the FDB-DS were utilised in this price-only experiment. Each ticker has approximately an average of 30,000 per minute share prices associated with it. These share prices were originally extracted in Section 5.3.2.

6.5.1 Methodology

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

1. Firstly, decide time periods used for this experiment, i.e. 1 day, 3 days and 5 days. This means there will be nine calculations, i.e. “SMA 1 Day”, “SMA 3 Day”, “SMA 5 Day”, “WMA 1 Day”, “WMA 3 Day”, “WMA 5 Day”, “EMA 1 Day”, “EMA 3 Day” and “EMA 5 Day”.
2. Programmatically calculate all the nine calculations of the SMA, WMA and EMA using their formulas as below:

Eq. 1. $$SMA = \frac{p_1 + p_2 + \cdots + p_n}{n}$$

where, *p* = price; *n* = time period.

Eq. 2. $$WMA = \frac{(P*n + P(1)*n-1+\cdots P(n-1)*1)}{\left(n*\frac{(n+1)}{2}\right)}$$

where, *p* = price; *n* = time period.

Eq. 3. $$EMA_{[today]} = (P_{[today]}\ x\ K) + (EMA_{[yesterday]}\ x\ (1 - K))$$

where, *p* = price; *K* = 2 / (n + 1); *n* = time period.

3. Record all the nine calculation outcomes in the database table in the FDB-DS.
4. Next, calculate 5%, 10% and 15% increment on top of the nine calculations, which act as the thresholds. As discussed in Section 5.8.1.1, for example, if a

particular “SMA 1 Day” price is $15.4, the following is an example of the threshold calculations:

Table 6.6 Moving Average Threshold Price Calculation Example

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

5. Save the outcomes of the threshold calculations for all the nine moving average calculations in the FDB-DS.
6. Once all the nine moving average calculations and the threshold calculations are successfully performed, check each original price against its threshold figures. If any of the original prices exceeds any of the moving average thresholds, label it with “5%”, “10%” or “15%” thresholds.
    - “5%” means when an original price is 5% above the moving average value.
    - “10%” means when an original price is 10% above the moving average value.
    - “15%” means when an original price is 15% above the moving average value.
7. Calculate the total number of flagged prices that trigger the thresholds of “5%”, “10%” and “15%” for all the nine calculations by using the following sample of MySQL query:

```
SELECT sma_alert, COUNT(*)
FROM price_sma_1day
WHERE sma_alert IS NOT NULL
GROUP BY sma_alert * 1 ASC;
```

6.5.2 Results and Discussions

The results of all the nine calculations will be discussed in the following Section 6.5.2.1 for SMA, Section 6.5.2.2 for WMA and Section 6.5.2.3 for EMA.

6.5.2.1 Results (SMA)

Table 6.7 and Figure 6.4 represent the SMA results for the total number of flagged prices that triggered the “5%”, “10%” and “15%” moving average thresholds, for the time periods of 1 day, 3 days and 5 days respectively.

As shown in the results, “5%” threshold has the highest count of flagged prices across the three time periods, i.e. 1 Day, 3 Days and 5 Days. The result shows that the more time period involved in the moving average calculations, the more flagged prices in each threshold.

Table 6.7 Total Number of Flagged Prices that Exceeded Thresholds (SMA)

	1 Day	3 Days	5 Days
5%	449,121	855,214	1,081,339
10%	107,608	237,403	335,837
15%	104,500	242,169	348,978

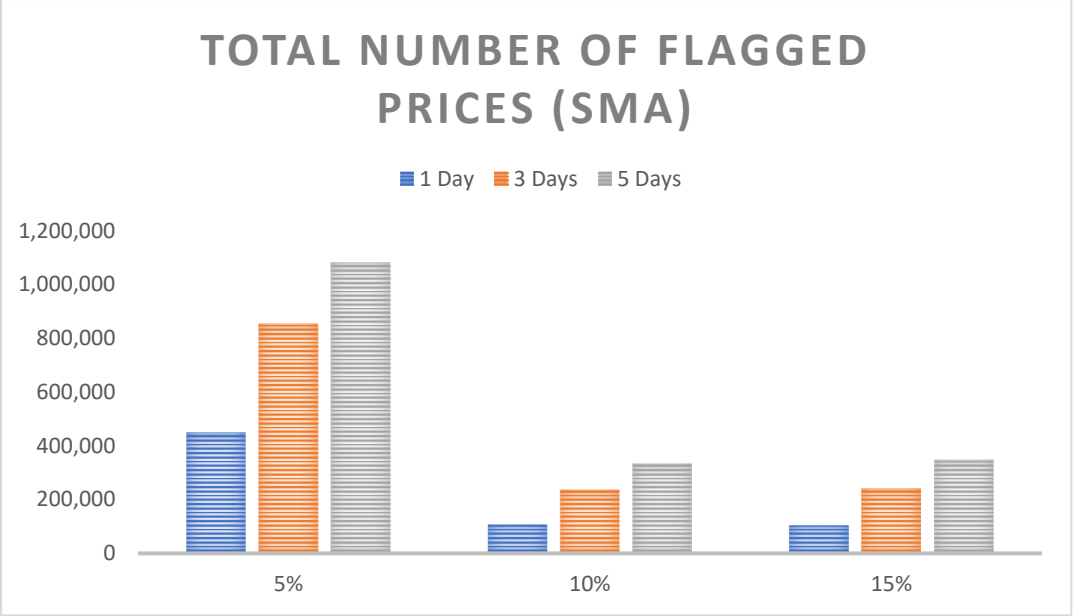

Figure 6.4 Total Number of Flagged Prices (SMA)

6.5.2.2 Result (WMA)

Table 6.8 and Figure 6.5 demonstrate the WMA results for the total number of flagged prices that triggered the thresholds of "5%", "10%" and "15%", for the periods of 1 day, 3 days and 5 days respectively.

The results appear to be similar to SMA; the "5%" threshold has the highest count of flagged prices across all the three time periods, i.e. 1 Day, 3 Days and 5 Days. The result shows that the more time periods involved in the moving average calculations, the more flagged prices in each threshold.

Table 6.8 Total Number of Flagged Prices that Exceeded Thresholds (WMA)

	1 Day	3 Days	5 Days
5%	227,578	609,372	812,138
10%	65,456	148,211	225,480
15%	74,163	151,284	210,320

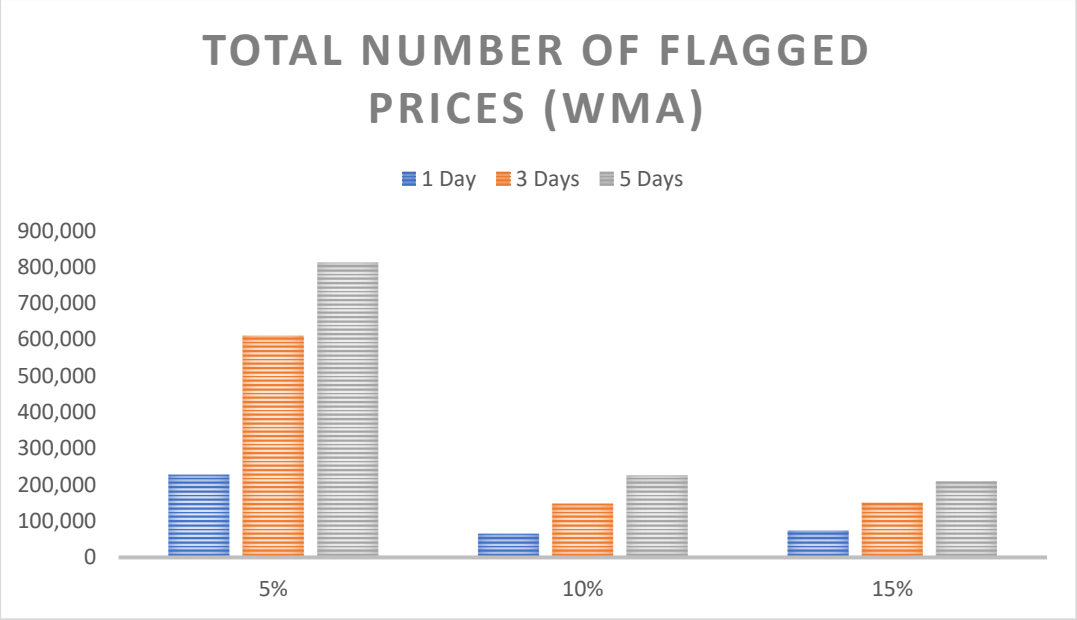

Figure 6.5 Total Number of Flagged Prices (WMA)

6.5.2.3 Result (EMA)

Table 6.9 and Figure 6.6 show the EMA results for the total number of flagged prices that triggered the thresholds for "5%", "10%" and "15%", for the time periods of 1 day, 3 days and 5 days respectively.

The results also appear similar to those for SMA and WMA; the "5%" threshold has the highest count of flagged prices across all three time periods, i.e. 1 Day, 3 Days and 5 Days. The result shows that the more days involved in the moving average calculations, the more flagged prices in each threshold.

Table 6.9 Total Number of Flagged Prices that Exceeded Thresholds (EMA)

	1 Day	3 Days	5 Days
5%	922,068	1,186,603	1,238,609
10%	396,272	431,148	513,299
15%	502,274	668,947	835,642

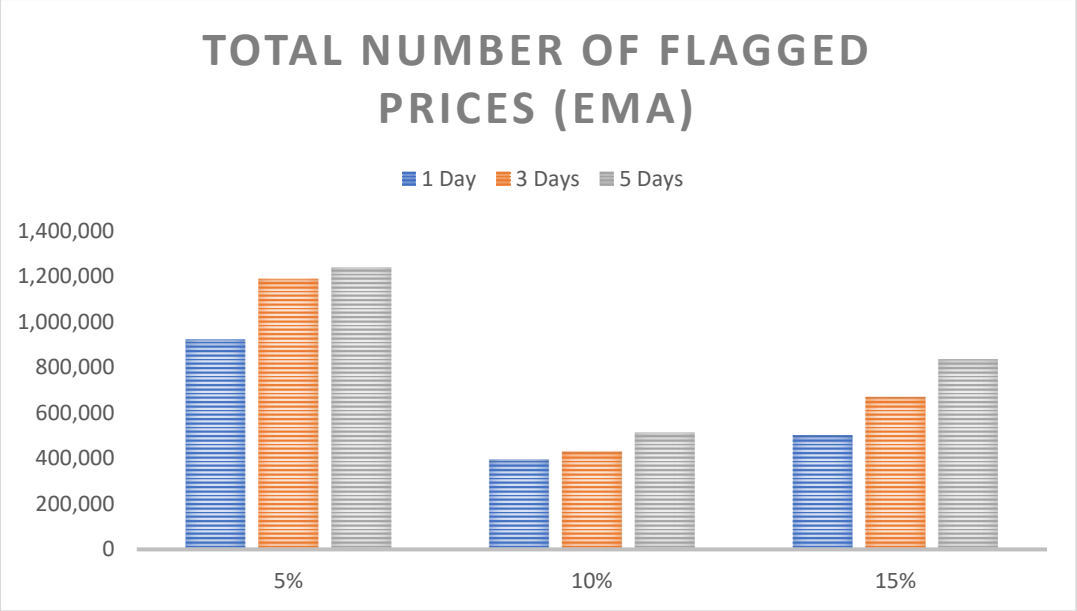

Figure 6.6 Total Number of Flagged Prices (EMA)

Moving average techniques can be used to flag the abnormalities in the price movements, which supports hypothesis H_{1c}. Based on observations, as compared to SMA and EMA, the prices flagged based on the WMA threshold calculations contain the least amount of flagged prices across all time periods of 1 Day, 3 Day and 5 Day. As mentioned earlier, the more time periods are involved, the more prices are flagged. EMA appears to have the most flagged prices of all. The flagged prices for SMA appears to be between WMA and EMA, which could potentially mean it is the better moving average to use as it is not leaning towards either too many or too few flagged prices. However, by flagging prices alone without looking at the flagged comment it does not tell a lot about whether the price hikes have any relation to the potential illegal comments. This leads to the establishment of Experiment 4 to backward label the moving average threshold to the flagged comments, which will be described in the next section.

6.6 Experiment 4: Backward Analysis of Prices and FDB Comments

The aim of this experiment is to apply the moving average thresholds calculated in Experiment 3 backwards to the flagged comments. This allows the further classification of the flagged comments in an attempt to reduce false positives for flagged comments.

This experiment tests the following hypothesis:

H_{0d}: Backward analysis cannot be performed by matching abnormal stock prices with the flagged comments to further classify flagged comments for reducing false positives.

H_{1d}: Backward analysis can be performed by matching abnormal stock prices with the flagged comments to further classify flagged comments for reducing false positives.

6.6.1 Methodology

The following describes the steps taken to perform the experiment:

1. Using the results from Experiment 3 (price-only experiment), the flagged prices that triggered the thresholds of “5%”, “10%” and “15%” for all nine calculations (i.e. “SMA 1 Day”, “SMA 3 Day”, “SMA 5 Day”, “WMA 1 Day”, “WMA 3 Day”, “WMA 5 Day”, “EMA 1 Day”, “EMA 3 Day” and “EMA 5 Day”) will be matched to the flagged comments by the dates of both price and comment.
2. For all nine calculations, label each flagged comment with the moving average thresholds of “5%”, “10%” or “15%” if it shares the same or nearest date and time as the flagged prices. Flagged comments that do not qualify for any moving average threshold labels will stay intact. Each label is explained as follows:
    - No moving average threshold labelling is needed if a flagged comment does not match any moving average thresholds.

- For all nine calculations, the label “5%” is applied to the flagged comment if the price that shares the same or nearest date and time is found to have exceeded the “5%” threshold.
- For all nine calculations, the label “10%” is applied to the flagged comment if the price that shares the same or nearest date and time is found to have exceeded the “10%” threshold.
- For all nine calculations, the label “15%” is applied to the flagged comment if the price that shares the same or nearest date and time is found to have exceeded the “15%” threshold.

3. Calculate the total number of flagged comments with all levels of price hike thresholds (“C”, “Y”, “A” and “R”) that matches the moving average thresholds (“5%”, “10%” and “15%”) for all nine calculations (i.e. “SMA 1 Day”, “SMA 3 Day”, “SMA 5 Day”, “WMA 1 Day”, “WMA 3 Day”, “WMA 5 Day”, “EMA 1 Day”, “EMA 3 Day” and “EMA 5 Day”).

6.6.2 Results and Discussions

In this section, SMA results will be discussed in Section 6.6.2.1, WMA results will be discussed in Section 6.6.2.2 and EMA results will be discussed in Section 6.6.2.3.

6.6.2.1 Results (SMA)

Table 6.10 and Figure 6.7 represent the total number of flagged comments that triggered both price hike thresholds (“C”, “Y”, “A” and “R”) and Simple Moving Average (SMA) thresholds (“5%”, “10%” and “15%”).

Table 6.10 Total Number of Flagged Comments that Triggered Both Price Hike Thresholds and Simple Moving Average (SMA) Thresholds

	5%			10%			15%		
	1D	**3D**	**5D**	**1D**	**3D**	**5D**	**1D**	**3D**	**5D**
C	518	1039	1300	204	291	366	242	356	395
Y	228	306	274	99	62	100	74	127	146
A	89	259	183	40	49	64	42	65	94
R	154	126	84	79	85	97	199	408	500

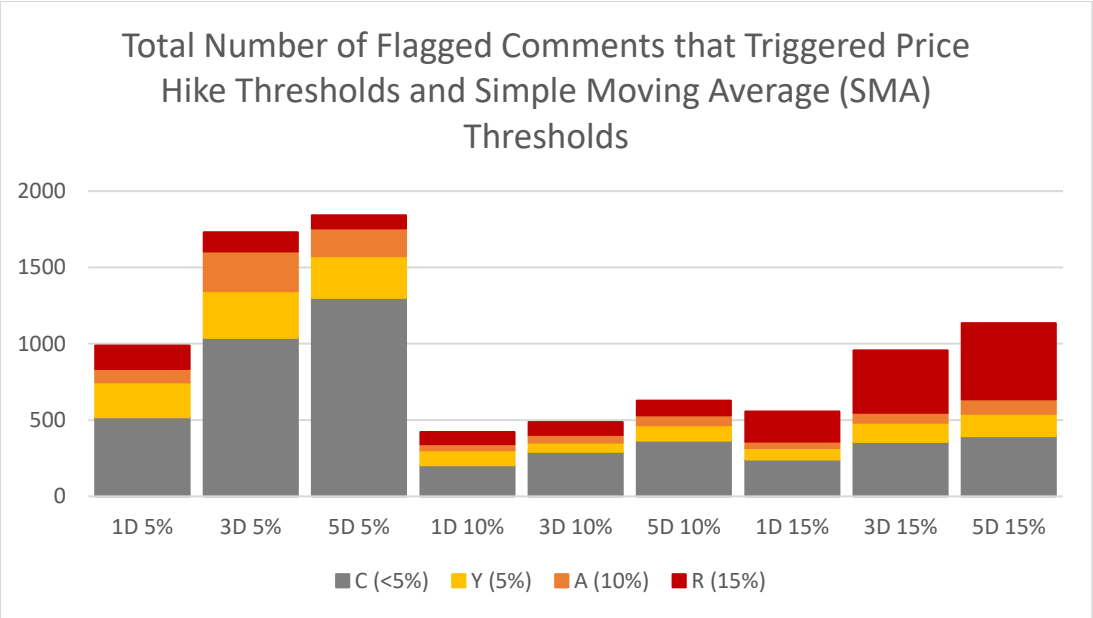

Figure 6.7 Total Number of Flagged Comments that Triggered Price Hike Thresholds and Simple Moving Average (SMA) Thresholds

In the table and chart, it can be seen that there are 228 flagged comments flagged as "Y" (5%) price hike threshold (from the forward analysis) and also flagged as "5%" for the "SMA 1 Day" moving average threshold (from the backward analysis). Then, 306 flagged comments are flagged as "Y" price hike threshold and "5%" moving average threshold for the "SMA 3 Day"; and, 274 flagged comments are flagged as "Y" price hike threshold and "5%" moving average threshold for the "SMA 5 Day".

As for the flagged comments that are labelled with "A" (10%) price hike threshold and "10%" moving average threshold, they are, in total, 40, 49 and 64 comments, for "SMA 1 Day", "SMA 3 Day" and "SMA 5 Day", respectively. Also, for the flagged comments that are labelled with "R" (15%) price hike threshold and "15%" moving average threshold, there is a total of 199, 408 and 500 comments, for the "SMA 1 Day", "SMA 3 Day" and "SMA 5 Day" respectively. It appears that there are more flagged comments under the threshold combinations of "Y" and "5%", and "R" and "15%". A user of the FDBM prototype system can prioritise these flagged comments before investigating other flagged comments.

Below is an example of the flagged comment that is being flagged with "Y" and "5%" for "SMA 1 Day", posted by the comment author "*TMPocket*" for the ticker symbol "BLU" (company name: Blue Star Capital) which is under the FTSE AIM All-Share index. This comment author was potentially trying to pump the share price by telling the FDB community that this is "hot stock".

> "*This will be a hot stock in the next 7 trading days and beyond the IPO will be very good for BLU shareholders and traders alike and will dwarf the price we are at now some fun and games to be had here but don't be caught out as this is serious multi bagger potential and a great company to be part of as a medium hold with trading ops. GL all."*

Below are two examples of the flagged comment that is being flagged with "R" and "15%" for "SMA 5 Day", posted by the comment author of "*systematic92*" for the ticker symbol "CRND" (company name: Central Rand Gold Ltd) which is also a listed company under the FTSE AIM All-Share index. The comment author simply convinced people by posting convincing words like "*told you*" and "*this is going up*", and even saying "*I'll ramp when the time is right*" which is a clear indication of the author committing a P&D crime.

> "*Told you monkeys. It's the Americans waking up too. This is going up again*"
>
> "*I'll ramp when the time is right. For now I will just say 11.67 baby*"

6.6.2.2 Results (WMA)

Table 6.11 summarises the total number of flagged comments that triggered both price hike thresholds ("C", "Y", "A" and "R") and Weighted Moving Average (WMA) thresholds ("5%", "10%" and "15%"). Figure 6.8 visualises the data in Table 6.11.

Table 6.11 Total Number of Flagged Comments that Triggered Both Price Hike Thresholds and Weighted Moving Average (WMA) Thresholds

	5%			10%			15%		
	1D	**3D**	**5D**	**1D**	**3D**	**5D**	**1D**	**3D**	**5D**
C	329	644	963	184	206	253	210	311	336
Y	174	258	288	79	83	66	70	99	115
A	64	110	191	33	37	40	33	53	68
R	157	126	107	58	78	99	145	315	392

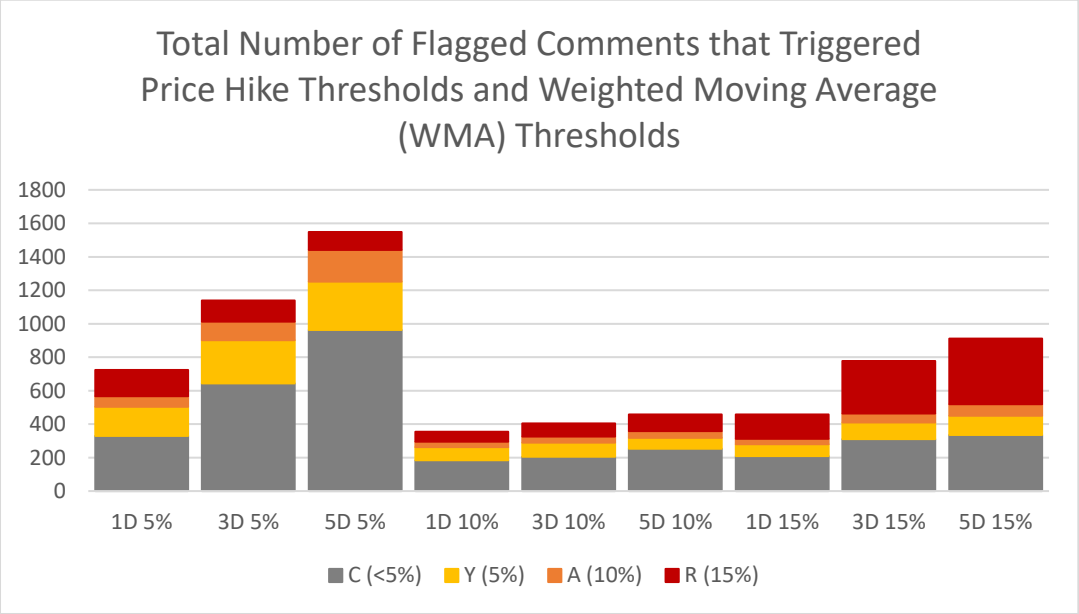

Figure 6.8 Total Number of Flagged Comments that Triggered Price Hike Thresholds and Weighted Moving Average (WMA) Thresholds

As depicted in Table 6.11 and Figure 6.8, there are 174 flagged comments being flagged by both "Y" (5%) price hike threshold and "5%" for the "WMA 1 Day" moving average threshold. There are 258 flagged comments that are being flagged as "Y" price hike threshold and "5%" moving average threshold for the "WMA 3 Day"; and, 288 flagged comments that are being flagged as "Y" price hike threshold and "5%" moving average threshold for the "WMA 5 Day".

As for the flagged comments that are labelled with "A" (10%) price hike threshold and "10%" moving average threshold, they are, in total, 33, 37 and 40 comments, for "WMA 1 Day", "WMA 3 Day" and "WMA 5 Day", respectively. Like the results in the SMA section, this threshold seems to contain a lower number of flagged comments. Next, for the flagged comments that are labelled with "R" (15%) price hike threshold and "15%" moving average threshold, there is a total of 145, 315 and 392 comments, for the "WMA 1 Day", "WMA 3 Day" and "WMA 5 Day" respectively. Similarly to SMA, it seems that there are more flagged comments under the threshold combinations of "Y" and "5%", and "R" and "15%".

Here is an example of the flagged comment that is being flagged with "A" and "10%" for "WMA 5 Day", posted by the comment author of "*Exptrader*" for the ticker symbol "COP" (company name: Circle Oil) which is under the FTSE AIM All-Share index. The comment author was telling the FDB community that the stock is "*about to pop*" which is a clear intention of trying to drive the price up.

"*Relentless buy now. This is about to pop. News is spreading.....gla*"

6.6.2.3 Results (EMA)

Table 6.12 and Figure 6.9 present the total number of flagged comments that triggered both price hike thresholds ("C", "Y", "A" and "R") and Exponential Moving Average (EMA) thresholds ("5%", "10%" and "15%").

Table 6.12 Total Number of Flagged Comments that Triggered Both Price Hike Thresholds and Exponential Moving Average (EMA) Thresholds

	5%			10%			15%		
	1D	**3D**	**5D**	**1D**	**3D**	**5D**	**1D**	**3D**	**5D**
C	904	2641	2765	236	269	467	427	576	682
Y	178	302	150	42	70	125	173	202	188
A	85	88	40	36	32	26	109	140	166
R	127	82	117	64	69	55	227	298	323

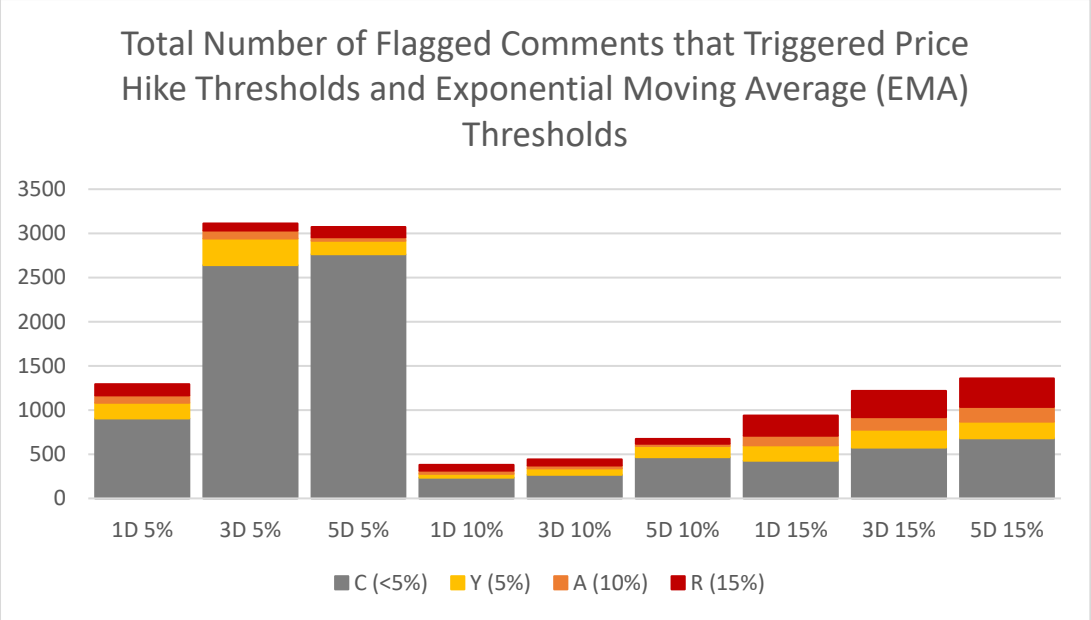

Figure 6.9 Total Number of Flagged Comments that Triggered Price Hike Thresholds and Exponential Moving Average (EMA) Thresholds

As shown in Table 6.12, there is a total of 178 flagged comments that were being flagged as "Y" (5%) price hike threshold and also flagged as "5%" for the "EMA 1 Day" moving average threshold. Overall, 302 flagged comments are being flagged as "Y" price hike threshold and "5%" moving average threshold for the "EMA 3 Day" and 150 flagged comments are being flagged as "Y" price hike threshold and "5%" moving average threshold for the "EMA 5 Day".

As for the flagged comments that are labelled with "A" (10%) price hike threshold and "10%" moving average threshold, they are, in total, 36, 32 and 26 comments, for "EMA 1 Day", "EMA 3 Day" and "EMA 5 Day", respectively. Also, for the flagged comments that are labelled with "R" (15%) price hike threshold and "15%" moving average threshold, there is a total of 227, 298 and 323 comments, for the "EMA 1 Day", "EMA 3 Day" and "EMA 5 Day" respectively. Like the results in the SMA and WMA sections, it appears that there are more flagged comments under the threshold combinations of "Y" and "5%", and "R" and "15%".

Below is an example of a flagged comment that is being flagged with "R" and "15%" for "EMA 5 Day"; the comment was posted by the author of "*DontBelieveHype*" for the ticker symbol "BOX" (company name: Boxhill Technologies) which is also under

the FTSE AIM All-Share index. In this flagged comment example, the comment author was trying to advise people not to follow and become the victims of P&D crime for the ticker symbol “BOX”.

> *“The rise is not based on anything else than speculation and probably 1000 posts an hour in a blatant attempt to move the stock which if I am not misguided is also known as market manipulation? I'm purely warning people here so they don't lose money from false promises.”*

The results presented in Section 6.6.2.1, Section 6.6.2.2 and Section 6.6.2.3 show that it is possible to perform backward analysis by matching the abnormal stock prices with the flagged comments to further classify flagged comments to resolve false positives. Relevant authorities or FDB moderators can indeed flag and detect potentially illegal activities happening on FDBs. Thus, this supports hypothesis H_{1d} in this experiment.

6.7 Statistical Test and Summary of Backward Analysis Experiments

The outcomes of both Experiments 3 and 4 appear to support the hypotheses. The data produced through the experiments are further validated through statistical tests.

The statistical tests for the backward analysis experiments were conducted using two separate methods. The first method was based on the assumption that the only important alerts (i.e. alerts that are worth focusing on) were the price jump to 5% or higher. The three alerts “Y” (5% price hike), “A” (10% price hike) and “R” (15% price hike) are recoded in SPSS prior to the tests with a value “1” and all other thresholds, i.e. “C” (less than 5% price hike) and “Null” are recoded with a value “0”. Similarly, the “5%”, “10%” and “15%” moving average thresholds are also recoded with a value “1” and recoded thresholds below 5% with a value “0”.

The variable “threshold” was renamed “Threshold_Yes_No” for better readability due to the recoding in previous steps. The variable “sma_1day_alert” was also renamed to “SMA1_Threshold_Yes_No”.

In this discussion, “SMA1_Threshold_Yes_No” is used for explanation. A crosstabulation table was constructed as shown in Table 6.13 and a chi-square test for independence was also conducted as shown in Table 6.14.

The other crosstabulation, chi-square and symmetric measure results that compare the other moving averages thresholds against “Threshold_Yes_No” are shown in Appendix F.

Table 6.13 Threshold_Yes_No * SMA1_Threshold_YesNo Crosstabulation – First Method

		SMA1_Threshold_YesNo		
		0	1	Total
Threshold_Yes_No	0	37,529	964	38,493
	1	10,361	1,004	11,365
Total		47,890	1,968	49,858

The crosstabulation data in Table 6.13 show that out of the total 49,858 cases, 37,529 cases fall into the non-important category in both “Threshold_Yes_No” and “SMA1_Threshold_YesNo”. But from there, there are a total of 1,004 cases which fall into the important category in both “Threshold_Yes_No” and “SMA1_Threshold_YesNo”. This means 1,004 of the flagged comments should be investigated for potentially illegal activities on FDBs at a higher priority.

Table 6.14 Chi-Square Test – First Method

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	927.245[a]	1	.000		
Continuity Correction[b]	925.576	1	.000		
Likelihood Ratio	777.525	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	927.226	1	.000		
N of Valid Cases	49858				

A chi-square test was conducted to examine the independence of the two variables i.e. “threshold” and “SMA 1 Day”. The test resulted in rejecting the null hypothesis (chi-square = 927.245, df=1, p-value <0.01). The correlation (r=0.136) between these two variables is also computed in Table 6.15 below and found to be significant.

Table 6.15 Symmetric Measures – First Method

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.136	.005	30.737	.000[c]
Ordinal by Ordinal	Spearman Correlation	.136	.005	30.737	.000[c]
N of Valid Cases		49858			

This means that the two variables are dependent and that the hypotheses in Experiments 3 and 4 appeared to be statistically supported.

Similarly, the “Threshold_Yes_No” variable is compared with the other moving average and time period i.e. “SMA 3 Day”, “SMA 5 Day”, “WMA 1 Day”, “WMA 3 Day”, “WMA 5 Day”, and “EMA 1 Day”, “EMA 3 Day” and “EMA 5 Day”. These results are shown in Appendix G. All the results show that the p-value of their chi-square tests is small (i.e. p-value<0.01) and hence, the null hypothesis of no association between the two variables is rejected. In another word, the hypotheses in Experiment 3 (H_{1c}) and 4 (H_{1d}) are accepted.

Table 6.16 Summary of Correlations – First Method

		sma1	sma3	sma5	wma1	wma3	sma5	ema1	ema3	ema5
Threshold	Co	.136**	**.150****	.133**	.128**	.143**	.143**	.096**	.032**	.004
	Si	.000	.000	.000	.000	.000	.000	.000	.000	.354
	N	49858	49858	49858	49858	49858	49858	49858	49858	49858

All hypothesis tests rejected the null hypothesis of no association between the threshold variable and moving average variables. All the correlations were positive and found to be statistically significant even though they were not leaning towards the stronger side. Correlation results in Table 6.16 show that the highest correlation is between “threshold” and "sma3” (i.e. SMA 3 Day) alerts. Among SMA, WMA and EMA, EMA seems to have the lowest correlations. As with the observation of the outcome of Experiment 3 (price-only experiment), SMA seems to be a more suitable moving average technique to be used with the backward analysis. In this case, the relevant authorities could consider using just the SMA calculations (as part of the backward analysis) when running investigations.

The second method of testing the data in the backward analysis experiments is described next. The results of the statistical tests in relation to the “threshold” and “SMA 1 Day” are discussed. This test, which slightly differs from the first method, is based on the assumption that all the thresholds “Y” (5% of price hike), “A” (10% of price hike) and “R” (15% of price hike) are important alerts and should be tested separately. Hence, “Y” is recoded as value “1”, “A” is recoded as value “2”, “R” is recoded as value “3”, “C” and “Null” are recoded as value “0”. This makes the “threshold” variable ordinal data, which is in order instead of it simply being categorical as in the first test. Similarly, the thresholds “<5%”, “5%”, “10%” and “15%” in moving average price calculations are also recoded as values “0”, “1”, “2” and “3” respectively.

In this discussion, “SMA 1 Day” is used for the explanation. A crosstabulation table was constructed as shown in Table 6.17 and a chi-square test for independence was also conducted as shown in Table 6.18.

The other crosstabulation, chi-square and symmetric measure results that compare the other moving averages thresholds against "threshold" are shown in Appendix G.

Table 6.17 Threshold_Recoded * SMA1_Threshold Crosstabulation – Second Method

		SMA1_Threshold				
		< 5%	5%	10%	15%	Total
Threshold_Recoded	C/N	37,529	518	204	242	38,493
	Y	4,796	228	99	74	5,197
	A	2,384	89	40	42	2,555
	R	3,181	154	79	199	3,613
Total		47,890	989	422	557	49,858

Table 6.18 Chi-Square Test – Second Method

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1319.065[a]	9	.000
Likelihood Ratio	932.526	9	.000
N of Valid Cases	49858		

Table 6.19 Symmetric Measures – Second Method

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.146	.007	32.849	.000[c]
Ordinal by Ordinal	Spearman Correlation	.141	.006	31.843	.000[c]
N of Valid Cases		49858			

Table 6.20 Summary of Correlations – Second Method

		sma1	sma3	sma5	wma1	wma3	sma5	ema1	ema3	ema5
Threshold	Co	.146**	.193**	**.202****	.125**	.171**	.191**	.122**	.096**	.084
	Si	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	49858	49858	49858	49858	49858	49858	49858	49858	49858

All hypothesis tests in the second method also rejected the null hypothesis of no association between the threshold variable and moving average variables. All the correlations were positive and found to be statistically significant even though they were not leaning towards the stronger side. However, the chi-square test value and correlation values were found to be improved when each threshold level is treated differently. This means that it makes sense to take into account separately the “Null” and “C”, “Y”, “A” and “R” as well as the “<5%”, “5%”, “10%” and “15%” on the moving averages side. Correlation results in Table 6.20 show that the highest correlation in the second method is between “threshold” and "sma5” (i.e. SMA 5 Day) alerts. Similar to the first method of testing, among SMA, WMA and EMA, EMA seems to generate the lowest correlations. Again, in this second method, SMA has once again hit the highest correlation value. It is, thus, synchronised with the observation in the outcome of Experiment 3 (price-only experiment) where SMA was shown to be a more suitable moving average technique to be used with the backward analysis because it does not lean towards too few or too many flagged prices. In this case, the relevant authorities could consider using just the SMA calculations (as part of the backward analysis) when performing investigations.

6.8 Chapter Summary

The purpose of Chapter 6 was to conduct four experiments in order to test the two novel methodologies, namely forward analysis and backward analysis, in relation to the research question. Two experiments belonged to forward analysis and another two experiments belonged to backward analysis.

The aim of the forward analysis is to flag and filter potentially illegal Pump and Dump (P&D) comments. This analysis flags the comments against the predefined P&D Information Extraction (IE) keyword template and then calculates the ±2 days’ worth of per minute share prices against the “base price” of the flagged comments. The flagged comments are then labelled with price hike thresholds (i.e. “C”, “Y”, “A” and “R”) accordingly. This allows the flagged comments to be filtered according to their price movements and the indices they belong to. The hypotheses for the two

experiments under forward analysis have both been supported empirically and statistically.

As for the backward analysis, the aim is to detect abnormalities in the price movements followed by performing backward analysis to match the abnormal stock prices with the flagged comments to further classify flagged comments with the intention of reducing false positives for flagged comments. This analysis flags the abnormalities in the price movements using moving average techniques (i.e. SMA, WMA, EMA) with the moving average thresholds (i.e. “5%”, “10%” and “15%”) for three different time periods (i.e. 1 Day, 3 Day and 5 Day) and then backward labels the moving average thresholds to the flagged comments, which attempts to further filter the flagged comments while resolving false positives. The hypotheses for the two backward analysis experiments have both been supported empirically and statistically.

This concludes that it is possible to perform both novel forward and backward analysis by flagging the comments using the P&D IE keyword template with the price hike thresholds, then match the abnormalities in the stock prices with the flagged comments to further classify flagged comments to resolve false positives.

This research also investigates the potential use of Semantical Textual Similarity (STS) in order to test whether it can be used to improve the comments flagging process in the forward analysis, by comparing the semantic meaning between the comments and the P&D keywords, phrases and sentences. Chapter 6 consists of three experiments for this purpose.

Chapter 7. Semantic Similarity Measures for Analysis of Financial Discussion Board Comments

7.1 Introduction

The use of the forward analysis and backward analysis to establish a relationship between price fluctuations and potential Pump and Dump (P&D) activities on Financial Discussion Boards (FDBs) was shown to be useful for the relevant authorities, in order to perform the detection of potentially illegal P&D activities on the share price based FDBs. It allows the detection and flagging of comments based on a predefined P&D IE keyword template with the involvement of share price movements. Price hike thresholds and moving average thresholds allow the relevant authorities to focus on an investigation based on the levels of price abnormalities.

The aim of the development of the FDBM prototype system is to aid the relevant authorities to monitor the comment posting activities on the FDBs, hence, more analysis, in terms of textual format, can be explored. Since comments and the P&D Information Extraction (IE) keyword template are both in textual format, the degree of semantic similarity between the comments and the keywords, phrases and short sentences can be evaluated.

Semantic similarity refers to “How close do these two sentences come to meaning the same thing?” (O’Shea, 2013). The aim is to let the computer understand both the syntactic and semantic meaning of a sentence or short text in a similar way to a human being. As these measures are designed to deal with short texts, they are ideal for use on the comments from FDBs as the language used does not have to be well formed into sentences for the similarity to be calculated.

In this chapter, the Semantic Textual Similarity (STS) approach is used to test the similarities between the FDB comments and the predefined keywords, phrases and short sentences in the P&D IE keyword template. Section 7.2 gives an overview of the application of STS to FDB comments in this research and the experiment subsections follow on from this.

The main hypothesis to be tested in this STS chapter is as follows:

H_0: The use of Semantic Textual Similarity (STS) for template-based detection of the potential Pump and Dump activities cannot improve the detection accuracy of the FDBM prototype system.

H_1: The use of Semantic Textual Similarity (STS) for template-based detection of the potential pump and dump activities can improve the detection accuracy of the FDBM prototype system.

The main hypothesis can be broken down into a few testable hypotheses which are tested in two experiments in Section 7.3 and Section 7.4. Lastly, Section 7.5 concludes this chapter.

7.2 Application of STS to FDB Comments Overview

7.2.1 Methodology for Data Generation

7.2.1.1 Aim

The aim of this data generation is to create a separate dataset which is meant for use in the STS experiments in this chapter. This separate dataset is created based on selecting a number of the comments from the Financial Discussion Boards dataset (FDB-DS) formed in the previous chapter. As one of the experiments in this chapter involves a human expert, it is impossible for the human expert to read all the comments. Thus, only 32 rows of comments will be selected from the FDB-DS. Overall, 16 of them are the comments that were flagged by the forward analysis methodology through the FDBM prototype system, which can be picked from the `flaggedcomment` table in the FDB-DS. Another 16 comments are non-flagged comments, which can be picked from the full `comment` table in the FDB-DS. Furthermore, according to O'Shea et al. (2014), this is a big enough sample size to validate the use of the STS approach.

7.2.1.2 Methodology

This section describes the methodology to create the 32-row comments dataset for the human expert and STS experiments:

1. Firstly, programmatically and randomly choose 16 rows of flagged comments from the `flaggedcomment` table in the FDB-DS.
2. Next, also programmatically and randomly choose 16 rows of non-flagged comments from the `comment` table in the FDB-DS.
3. Merge both sets of 16 comments from Step 1 and 2 into an Excel spreadsheet.
4. Use a random sorting function in Excel; randomly sort all the 32 comments so that the flagged and non-flagged comment sequence is randomised.
5. Label all the randomised comments rows C1 to C32.

7.2.1.3 Dataset

Table 7.1 below is a result of the dataset generation with the 32 rows of comments randomly sorted and labelled from C1 to C32. In Table 7.1, the column "FDBM's Answer" represents the outcomes of the flagged comments and non-flagged comments. All the 16 flagged comments detected by the FDBM prototype system in the previous chapter are labelled as "y", whereas the non-flagged comments do not have any labels. This 32-row dataset will be used starting from Experiment 1 which involves a human expert labelling the comments that are deemed potentially illegal.

Table 7.1 32-row Comments Dataset

ID	Comment	FDBM's Answer
C1	Many of us have been around here for ages! It's been manipulated and diluted and now! Well it's a joke. There must be some sort of regulation even for aim	y
C2	That's champion that is. Wetherspoons…	
C3	derampers fam etc keep the price down for me for 2 more weeks I'll have 10 mill then lol	
C4	http://www.fca.org.uk/static/documents/short-positions-daily-update.xls	y

C5	brad fall down a well does anyone know?	
C6	Expanding into Namibia like expanding into empty space is no particular coup.The county's entire population of 2.1 million people is less than half that of Dar es Salaamalone.http://en.wikipedia.org/wiki/Namibia	
C7	i've been away for awhile - what's been going on here (apart from sp drop)?	y
C8	i cant see them they must be through isdx	
C9	Couldn.t resist. Another 10000 on Friday to add to my holdings and at 10.00p. Now come on Brad where is that RNS with all the lovely news of more drills happening more oil being pumped bank debt reduced. We know from experience what happens to the SP on good news. It flies and not at 1/10th penny a time. The long term holders must still be here and with limited stock available it can only go one way. After my last 'buy' I promised Mrs. Smidsy no more. I'm now on silent meals and sleeping in the spare room. LOL It won't be that when we get the RNS.	y
C10	Strange drop this morning on little volume unless we see some delayed sells? I've taken the opportunity of the dip in SP to top up substantially. Hoping for news on the processing plant and sales.	y
C11	Why are people posting RT sold out to the shorters in an attempt to stop the short attack? Utter utter rhubarb.	y
C12	Between lectures Wormfool? I am wondering why you aren't pointing out that xcite managed to get this oil to flow where others (including Chevron) couldn't do so? Ah yes I remember you are a boiler room deramper (part-time) to pay for your student digs. ATB	
C13	True! Communication is key. Today's rise was proof of that (though partly also a rebound from the quick drop it had). Will be interested to see what tomorrow brings.	y
C14	For the record I am long and have held firm thru the lows...it's just I haven't offered shorting advice that contradicts the position that I hold.	y
C15	Possibly. I think general concensus is that we shouldn't drop below high 3's again - so people may think when no 7/8am RNS (and an intra day fairly unlikely) that they can make a few quid intra day. Will / should always find support at this level. Wish I had the guts / skill / luck	y
C16	Blimey we are getting cheesed off here. I got in at 10.5p or so so not too bad but still a 50% plus drop. Yet I still have faith in the Board. From what we glean and extrapolate from info released things appear to be moving in the right direction. It's a slow and frustrating process. Maybe for the BoD too assuming they look at this site. Don't forget that they are	y

	shareholders as well and have a lot at stake. I don't think that they have lost enthusiasm but there must be a lot going on in the background. E.g. Visitors to the site. A 1million loan can't be made as a last throw of the dice surely? More likely with an expectation of it being repaid when something big happens. Is this wishful thinking? Maybe. Am I tempted to top up? yes but I think I will wait yet...GLA	
C17	It worked!!	
C18	Trend going on lately in aim market and difficult to call anything. I thought small miners were finished for now and look what happens a lot of them go crazy you just couldn't make it up especially when you look at their larger counter parts here I just don't know anything possible but these drops seem rather large considering some of the companies on offer but cash is king and the problem here I expect watching closely took a punt on Blu and prem could add this to the basket!	y
C19	Flow rates out Monday morning? Perhaps? Time to buy now before we bounce to 6p on flow rates Then TI should complete adding more $$$$ to our increasing monthly income. Just topped up another 260000 units will try and add more later once I've had a little shuffle round on stocks good 15-20% coming next week I feel. Gla.	y
C20	At 2p. ok I'm confused…	
C21	Agreed Bel. But I think SAV has different resources. SAV doesn t have iron. Secondly the price way down the placing price at 4.45p. So the drop is overdone. Massively oversold. Expect the bounce very soon.	y
C22	Thats a great article on HH. It explains it all about how irrational some investors are when it comes to news. I think the MMs know some investors would panic with no show of the black stuff & sell. They stopped short on the well so they would not have a blow out if they encountered oil. The presence of gas indicates there is oil. They saw it in the mud anyhow.	y
C23	Ask now @ 6	
C24	double code 1	
C25	Persimmon remains confident about market Persimmon said that it remained encouraged by the level of customer confidence in the UK housing market as it reported its progress from July 1 to November 3 2014 today (November 4). During the period which the volume housebuilder said reflected an expected return to a more traditional seasonal pattern to customer activityit achieved around 696 million of forward sales reserved beyond 2014 a 12% increase on last year. Its private sale reservation rates were around 2% lower than the same period last year. Persimmon pointed out the tough comparatives from 2013 when the government introduced the	

	Help to Buy equity loan scheme and its private reservations surged 45% ahead of 2012. Its visitor levels equalled those of last year. Persimmon also said that although it had successfully opened 80 of the 100 new sites scheduled for the second half of the year opening new sites without undue delay remained one of the industrys biggest constraints. It is currently selling homes from around 375 active development sites. But Persimmon stated that it was still confident it would deliver further growth in group operating profits in the second half. With our new site openings land recoveries continue to improve and with increased production volumes we are capturing further build efficiencies the companys statement read.	
C26	I surmise your talking about the one and only ? If so well done gl tick tock	
C27	Wow we have some right characters on tonight. Looking for minimum revenue of 28m for the 6 months positive cash flow and a small loss due to 1.5m restructuring costs. Not sure how anyone can de-ramp until results are seen but guessing they have their own agenda.	y
C28	listen.....fill yer boots it will be 37 again soon trust me	
C29	I'm not familiar with any of the branch workings as I work in the Asset Finance side our head office is located just behind the Cardiff Central train station. I do have a requirement to travel to Chiswel Street some weeks but if I'm completely honest London is just to busy for my liking.	
C30	Well Anth I reckon people are doing exactly what you're talking about doing without the ISA buyback. There must be some good gains out there. I'm guessing people are crytallising some gains offsetting them with a BOR loss and converting to cash before the end of December probably will prove to have been a good idea. I wish I had decent gains to make. POOM. If you've managed to not need glasses til 72 you are doing well. i needed them by 47.	
C31	LEEK or LEAK anyone ? D21	
C32	What would you suggest trying to explain basic economics to the readers of the Sun / Guardian you know the sort of thing the difference between deficit and debt? But hey ho I'm not holding stock (nor do I have any short positions either) not am I holding on to the mindless idea that one can talk the market down by 20% a day (and repeat the feat) by talking *******s.PS are you interested in my Unicorn farming venture I have three breeding pairs and plans to buy 5 more.	y

7.2.1.4 Graphical User Interface (GUI) for STS Experiments

The original STASIS (Li et al., 2006) algorithm Python script has been slightly modified in order to accommodate all the short sentences, phrases and artificially created sentences based on the keywords from the P&D IE keyword template. These short texts are used in Experiment 2 in Section 7.4 later. The purpose of adding a simple GUI for STASIS, as shown in Figure 7.1, is to make the experimental process easier, in terms of time consumption during the similarity computation between the short texts and the 32-row comments dataset. Another minor modification was also made to write all the similarity results into a CSV file, which also makes analysis easier in the experiment.

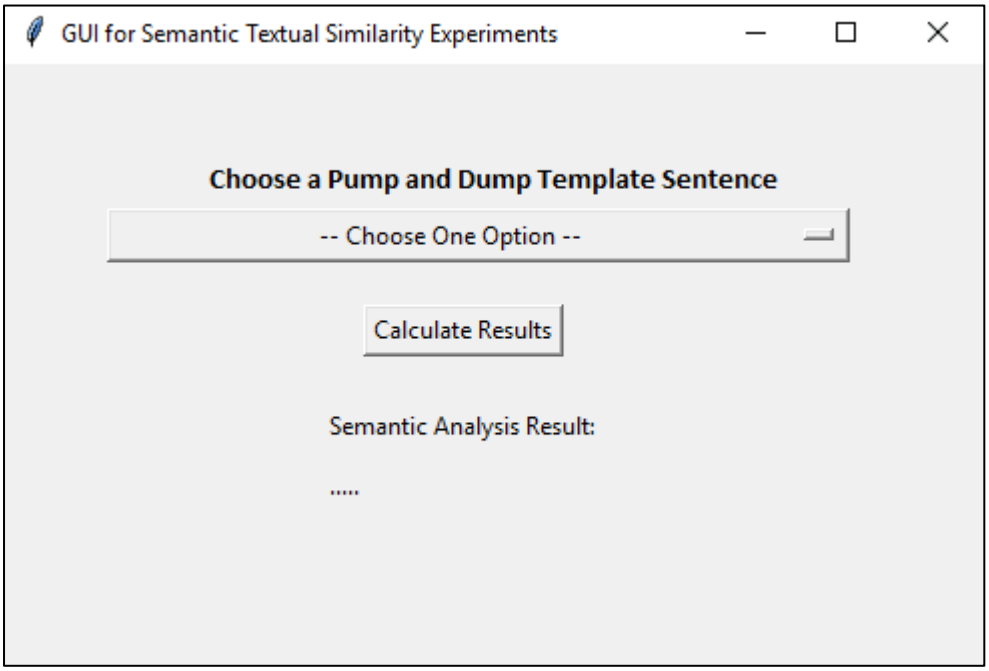

Figure 7.1 GUI for Semantic Textual Similarity Experiments

7.3 Experiment 1: Human Expert Comments Labelling

The aim of the experiment is to test whether a human expert in the field can determine the flagged comments and non-flagged comments correctly as per the answers according to the FDBM prototype system.

7.3.1 Hypothesis

The testable hypothesis for Experiment 1 is listed as follows:

H_{0b}: A human expert cannot determine the flagged comments and non-flagged comments correctly as per the answers according to the FDBM prototype system.

H_{1b}: A human expert can determine the flagged comments and non-flagged comments correctly as per the answers according to the FDBM prototype system.

7.3.2 Methodology

This section describes the methodology to conduct Experiment 1:

1. Without disclosing the FDBM's answer to the human expert, the human expert is asked to flag the 32 randomly sorted comments (produced in Section 7.2.1.3) that were seen to be indicative of potential P&D and non-potential P&D.
2. Once the human expert has labelled all the 32 rows of comments, compare the results of both FDBM's answer and expert's answer.

7.3.3 Results and Discussions

Table 7.2 below represents the answers from both the FDBM prototype system and the human expert.

Table 7.2 32-row Comments Dataset – With Human Expert's Answers

ID	FDBM's Answer	Expert's Answer
C1	y	y
C2		
C3		y

C4	y	
C5		
C6		
C7	y	
C8		
C9	y	y
C10	y	y
C11	y	y
C12		y
C13	y	
C14	y	
C15	y	y
C16	y	y
C17		
C18	y	
C19	y	y
C20		
C21	y	y
C22	y	y
C23		y
C24		
C25		
C26		
C27	y	y
C28		y
C29		
C30		
C31		y
C32	y	

Comparing the answers between those that are labelled by the FDBM prototype system and those that are labelled by the human expert, the human expert has labelled 10 flagged comments and 11 non-flagged comments which match with the results from the FDBM prototype system. This yields a 65.6% accuracy in terms of the matches between both results.

There are six flagged comments (determined by FDBM) that the human expert labelled as non-potentially illegal. There are also five non-flagged comments (determined by FDBM) that the human expert labelled as potentially illegal. To understand the discrepancies, the results are evaluated. From Table 7.1, take the

comment “C4” as an example (which appears to be just a link, i.e. “*http:// www.fca.org.uk/static/documents/short-positions-daily-update.xls*”).

FDBM labels comment “C4” as a potentially illegal comment, whereas the human expert did not label it as a potentially illegal comment. The entire comment in “C4” is merely a website link. FDBM picked it up as potentially illegal mainly due to the matching keyword. On the other hand, the human expert argued that the link pointing to the Financial Conduct Authority (FCA – a financial regulatory body in the UK which operates independently from the UK government) does not represent a potentially illegal comment. It is true that the link and the content of the link do not show any indication of an illegal comment. However, as discussed earlier in the experimental results in Chapter 6, from the FDBM prototype system’s perspective, if a comment is labelled as a potentially illegal comment, this will direct the relevant authority’s attention to allow them to decide whether to inspect the FDB discussions surrounding a particular flagged comment on the same day, three days or even five days.

7.4 Experiment 2: STS Approach for Comments Flagging

The aim of this experiment is to test whether the STS approach can be associated with the process of comments flagging in the forward analysis, in order to allow the relevant authorities to prioritise the investigation based on the STS similarity scores. For example, if the similarity score is higher than a certain threshold when comparing a keyword, phrase or short sentence to a comment, then this comment is worth investigation by the relevant authorities.

The P&D IE keyword template, presented in Section 5.6, contains the following 20 short sentences:

1. "When I post you better believe me"
2. "Do not doubt me on this one"
3. "Anyone who sells to make a small profit is very short sighted"
4. "This is the chance"

5. "Price will go up"
6. "Buy as quickly as possible"
7. "Take what profit you can"
8. "About to run, get in today"
9. "Adding a few more don’t let the shorts win here"
10. "Buying opportunity this AM"
11. "Still looks good for a run"
12. "This is the opportunity"
13. "shall go up again after today"
14. "Make a pretty penny"
15. "Now is the time"
16. "Chance to make some real dollars"
17. "This stock might not have any trouble hitting over"
18. "May be too late if we wait a bit longer"
19. "Sell as quickly as possible"
20. "Get out while you can"

The P&D IE keyword template also contains a total of 97 keywords and phrases, which the keywords are artificially transformed into longer texts so that they can be tested using the STS approach. The following list demonstrates a list of the artificially created very short texts and the original phrases from the P&D IE keyword template:

1. "Is this Pnd"
2. "Did he try to Manipulate"
3. "Are you Manipulating?"
4. "This is Misleading"
5. "Are you deceiving"
6. "It is a Hoax"
7. "It is a Scam"
8. "Are you Falsifying"
9. "Is that a Cheat"
10. "Are you Spreading rumours"
11. "Artificially raising price"
12. "He is Pumping"
13. "Are you Ramping"
14. "Is this a tip"
15. "Are you elevating"
16. "Are you boosting"
17. "Might Get rich"
18. "Make a killing"
19. "Make a fortune"
20. "Make money"

21. "Make big bucks"
22. "Make a bundle"
23. "Make a bomb"
24. "Make a packet"
25. "Hit it big"
26. "You will Gain profit"
27. "You can Become wealthy"
28. "Strike it rich"
29. "Dump it now"
30. "Is he Dumping"
31. "I am Shorting"
32. "Do not Opt out"
33. "Do not Abandon"
34. "Do not Back out"
35. "Give up the ship"
36. "Do not Pull out"
37. "Do not Let go"
38. "Pump and dump"
39. "pump dump"
40. "you are Manipulating stock"
41. "Chop stock"
42. "That is a False statement"
43. "Misleading statement"
44. "Misleading positive statement"
45. "Once-in-a-lifetime"
46. "Once in a lifetime"
47. "Pump the price"
48. "Pump the share"
49. "Pump this stock"
50. "it has an Inflated price"
51. "This is a Hot stock"
52. "Huge volume spike"
53. "Let us Keep ramping"
54. "On hypes"
55. "You should Buy now"
56. "It has a Good future"
57. "Invested so heavily"
58. "It will fly"
59. "You should Buy more"
60. "Rock bottom price"
61. "Buy on dips"
62. "These are the Best tips"
63. "It is a Good tip"
64. "Dump the price"
65. "Dump the share"
66. "It will Fall hard"
67. "Sell it now"
68. "You should Sell quickly"
69. "It is a Declining stock"

The list contains 69 short texts instead of 97 short texts because some keywords are redundant when they are transformed into short texts. For example, the keyword “manipulate”, “manipulation”, “manipulating” and “manipulative” are all transformed into just two very short texts instead of four very short texts.

7.4.1 Hypothesis

The testable hypothesis for this experiment is listed as follows:

H_{0b}: The STS approach cannot be used to flag comments by comparing the semantic meaning between the comments and the keywords, phrases and short sentences in the P&D IE keyword template.

H_{1b}: The STS method can be used to flag comments by comparing the semantic meaning between the comments and the keywords, phrases and short sentences in the P&D IE keyword template.

7.4.2 Methodology

This section describes the methodology to conduct the experiment:

1. Execute the GUI-enabled STASIS Python script (Section 7.2.1.4), which already has the 32 comments and the 89 short texts (which include short sentences, phrases and artificially created short text based on keywords in the P&D IE keyword template) preloaded.
2. Compute the semantic similarity between all the 32 comments and all the 89 short texts.
3. Record all the similarity results produced through the CSV file.
4. Set similarity thresholds of 0.5, 0.55, 0.6, 0.65 and 0.7 and empirically compare the STS results with the answers by a human expert (from Experiment 1) and FDBM. The semantic similarity threshold of 0.5 was chosen as the lowest point as this allows observations of moderate to higher similarity.

7.4.3 Results and Discussions

In Experiment 1, Table 7.2 has shown the potential illegal comment labelling performed by the human expert as compared to the ones that were flagged by the FDBM prototype system using the P&D IE keyword template. In this experiment, the semantic similarity between each comment and each short text are computed.

On the X-axis in Figure 7.2 and Figure 7.3 below each of the 32 rows of comments (i.e. C1 to C32) from Table 7.1 in the previous section are represented, whereas the Y-axis represents the total count for the short texts with similarity scores which are higher than the semantic similarity thresholds. On first glance, it can be seen in Figure 7.2 and Figure 7.3 that by using a semantic similarity threshold of 0.5, a total of 14 comments (i.e. C1, C5, C9, C10, C13, C14, C16, C17, C19, C22, C23, C25, C29, C32) are worth investigating for potentially illegal activities on FDBs.

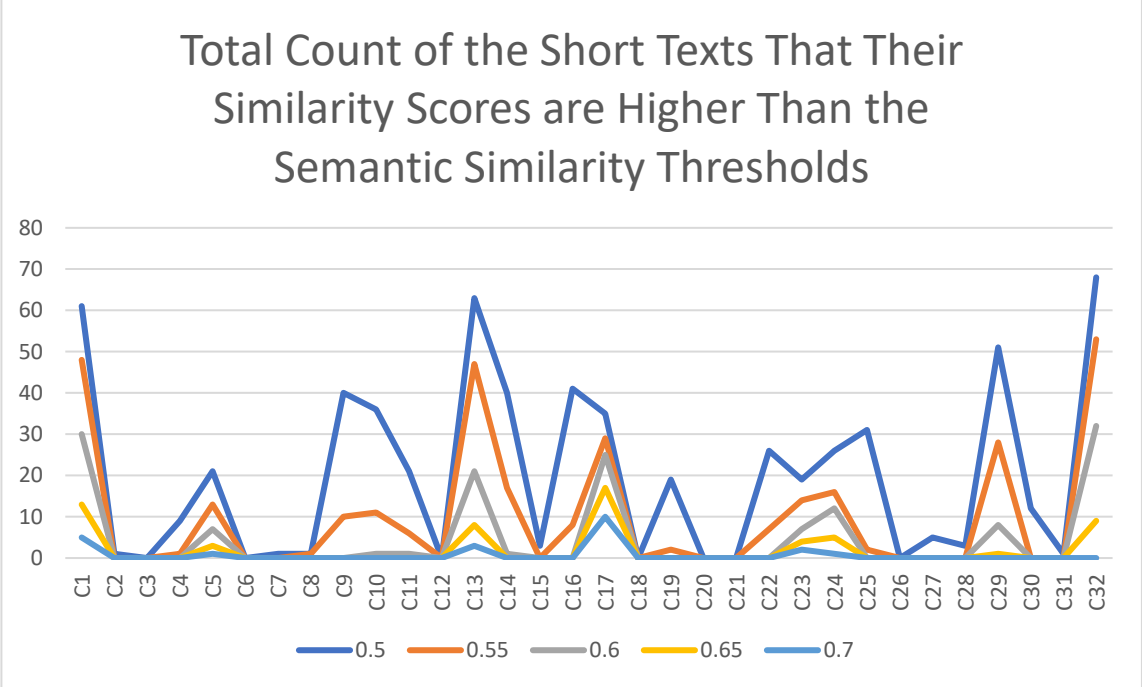

Figure 7.2 Total Count of the Short Texts whose Similarity Scores are Higher than the Semantic Similarity Thresholds (Line Chart)

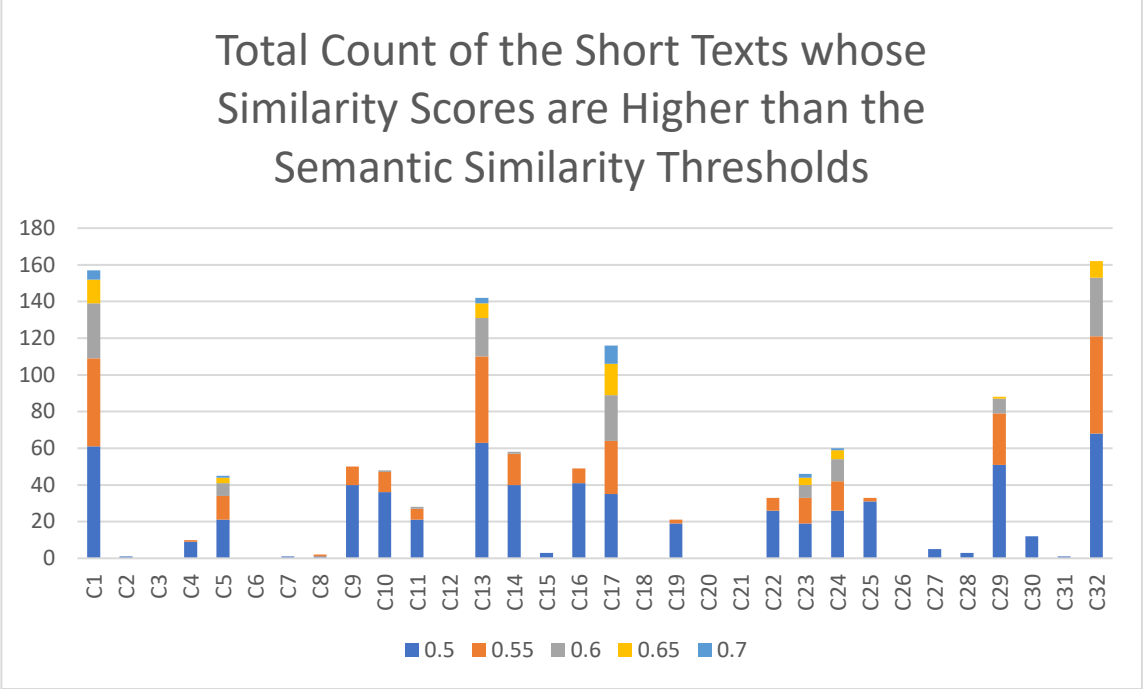

Figure 7.3 Total Count of the Short Texts whose Similarity Scores are Higher than the Semantic Similarity Thresholds (Stacked Bar Chart)

As can also be seen from Figure 7.2 and Figure 7.3, a semantic similarity threshold of 0.55 also shows an obvious pattern similar to the threshold of 0.5. The graph pattern for threshold 0.6 and above is not really obvious, in terms of flagging potentially illegal FDB comments.

Table 7.3 below represents the same data being visualised in Figure 7.2 and Figure 7.3, but with an additional two columns taken from Table 7.2 from Experiment 1 (i.e. the "FDBM's Answer" and the "Expert's Answer" columns). This allows the comparisons among the FDBM's Answer, Expert's Answer and the semantic similarity thresholds.

Table 7.3 Total Count of the Short Texts whose Similarity Scores are Higher than the Semantic Similarity Thresholds

ID	FDBM's Answer	Expert's Answer	0.5	0.55	0.6	0.65	0.7
C1	y	y	61	48	30	13	5

C2			1	0	0	0	0
C3		y	0	0	0	0	0
C4	y		9	1	0	0	0
C5			21	13	7	3	1
C6			0	0	0	0	0
C7	y		1	0	0	0	0
C8			1	1	0	0	0
C9	y	y	40	10	0	0	0
C10	y	y	36	11	1	0	0
C11	y	y	21	6	1	0	0
C12		y	0	0	0	0	0
C13	y		63	47	21	8	3
C14	y		40	17	1	0	0
C15	y	y	3	0	0	0	0
C16	y	y	41	8	0	0	0
C17			35	29	25	17	10
C18	y		0	0	0	0	0
C19	y	y	19	2	0	0	0
C20			0	0	0	0	0
C21	y	y	0	0	0	0	0
C22	y	y	26	7	0	0	0
C23		y	19	14	7	4	2
C24			26	16	12	5	1
C25			31	2	0	0	0
C26			0	0	0	0	0
C27	y	y	5	0	0	0	0
C28		y	3	0	0	0	0
C29			51	28	8	1	0
C30			12	0	0	0	0
C31		y	1	0	0	0	0
C32	y		68	53	32	9	0

It can be observed that, for example, comment C1 was flagged by the FDBM prototype system as potentially illegal; at the same time, it has a total count of 61 short texts whose similarity score is higher than the threshold of 0.5. In this case, it is assumed that the accuracy of the STS approach is matched by the FDBM's. Another example in Table 7.3 above is that comment C3 is not flagged as potentially illegal by FDBM and, at the same time, it has zero counts of short texts that exceed the similarity threshold of 0.5. In this

case, it is also assumed that the accuracy of the STS approach is matched by the FDBM analysis. Thus, due to these assumptions, there is a total of 19 cases (regardless of whether they are flagged or not) out of the 32 cases (i.e. the 32 comments) in Table 7.3 above that accurately match the STS approach and FDBM's analysis.

Looking deeper into the 19 cases (which contain both flagged and non-flagged comments), there are as many as 14 of them which are flagged comments. This means in Table 7.3 above there are a total of 16 flagged comments that were flagged by FDBM, and 14 of them are also flagged by the semantic threshold of 0.5. This yields an accuracy of 87.5%.

Next, by observing the semantic similarity threshold of 0.55, it can be seen that, when taking comment C15 for example, there is a zero count of short texts hitting the threshold of 0.55 despite having three counts for the threshold 0.5. In this case, it can be assumed that there is not an accurate match between the STS approach and FDBM's analysis for the threshold of 0.55. For the threshold of 0.55, it is observed that there is a total of 20 cases out of the 32 cases in Table 7.3 that are matched accurately between the STS approach and FDBM's analysis.

Looking deeper into the 20 cases (which contain both flagged and non-flagged comments), there are as many as 11 of them which are flagged comments. This means in Table 7.3 above there is a total of 16 flagged comments that were flagged by FDBM, and 11 of them are also flagged by the semantic threshold of 0.55. This gives an accuracy of 68.75%.

A semantic similarity threshold of 0.6 and above has less than 17 cases that are matched accurately between the STS approach and FDBM's analysis. The matched cases reduce in number when the threshold increases. It appears that the optimal threshold for the relevant authorities to use is 0.5 because it yields the highest accuracy of 87.5% when matching flagged comments with the short texts using the STS approach.

When comparing the STS approach to the expert's answers, for the threshold of 0.5 and 0.55, there are only 16 cases and 15 cases of accurate matches respectively, regardless

of flagged or non-flagged comments. For the threshold of 0.5, out of the 16 matched cases, there is a total of 12 flagged comments. Since there are 16 flagged comments in Table 7.3 above, this yields an accuracy of 75% because 12 out of the 16 flagged comments are correctly matched between the expert's answer and a similarity threshold of 0.5.

As for the threshold of 0.55, out of the 15 matched cases, there is a total of 8 flagged comments. Since there are 16 flagged comments in Table 7.3 above, this yields an accuracy of only 50% because only half (i.e. 8 flagged comments) out of the 16 flagged comments are correctly matched between the expert's answer and similarity threshold of 0.55.

The results and discussions in this experiment appear to reject the null hypothesis and accept the hypothesis (H_{1b}) as it is proven that it has the potential to aid the relevant authorities during the detection of potentially illegal FDB activities.

7.5 Chapter Summary

The purpose of this chapter was to conduct two experiments surrounding the Semantic Textual Similarity (STS) approach. The aim of these experiments was to test whether the STS approach can be adopted alongside the novel forward analysis and backward analysis. The STASIS algorithm (Li et al., 2006) has been used for experimental purposes because it takes function words and word orders into account in addition to its algorithm being more computationally efficient as compared to LSA. The results of the experiments were discussed and results suggest that the semantic similarity threshold of 0.5 is the optimal threshold for FDB comment data. This is due to the high accuracy being yielded from the experiments. For example, for the threshold of 0.5, out of all the 19 matched cases (a mixture of flagged and non-flagged comments), there is a total of 14 flagged comments flagged by both FDBM and the STS approach. This chapter concludes that STS has high future potential to be used alongside the novel forward and backward analysis in detecting fraudulent FDB posts.

Chapter 8. Conclusion and Further Work

8.1 Introduction

This chapter concludes the thesis by summarising each chapter, followed by describing the contributions of this research work in relation to the research aim and objectives. Future directions of this research work will also be discussed in the following sections.

8.2 Thesis Summary

This thesis presented a financial crime analysis methodology for Financial Discussion Boards (FDBs) using Information Extraction (IE) techniques, Moving Average (MA) techniques and a Semantic Textual Similarity (STS) approach.

Chapter 2 introduced the stock exchange in the UK, namely the London Stock Exchange (LSE, 2017). Then, it introduced the three share price based Financial Discussion Boards (FDBs) based in the UK whose share prices data are obtained from the LSE. However, the comments posted on these FDBs are not identical, unless some comment authors intentionally repeat the identical comments on all three FDBs, which is rare. While reviewing the three FDBs, the chapter also identified the semantically understandable artefacts on the FDBs. FDB artefact data like ticker symbol, comments and share prices were taken into account in experimental chapters. Subsequently, the stock market regulatory agencies in both the UK and the US were also presented. The chapter also highlighted the existing popular Pump and Dump (P&D) financial crime cases that have happened on FDBs which were mainly dealt with by the regulatory agency in the US. It then followed by giving reports of the existing research related to P&D financial crime which has happened through methods such as emails, social media and FDBs. Almost all the FDB research in the past has tended to focus on prediction through trading volume and quantified content from the Internet (Jin et al., 2016) instead of focusing on the share price movements and moderating the context of the content such as FDB comments. To the best knowledge of the thesis author, other than Delort et al. (2011)

who attempted to moderate the Online Discussion Sites (ODSs) using a machine learning approach, and the initial work presented by Knott and Owda (2012) using a knowledge engineering approach, there is no evidence of other work attempting to moderate FDBs using moderation tools.

Chapter 3 reviewed the three types of data structures that can be collected from the Internet. It then introduced the two fundamental classes of Information Extraction (IE), namely a knowledge engineering approach and automatic learning approach. A knowledge engineering approach, which is also known as a rule-based approach, has been discussed as it is the approach used by the template-based IE prototype system, namely FDBs Miner (FDBM) in this thesis. Semantic Textual Similarity (STS) was also introduced in this chapter.

Chapter 4 introduced a methodology (which comprised both novel forward analysis and backward analysis methodologies) and a novel architecture for the detection of potential financial crime activities on the share price based FDBs. The methodology for identifying the FDBs and the semantically understandable artefacts were described. This was followed by a description of an overall methodology (Section 4.3) for the development and implementation of the prototype system, which is divided into seven phases (i.e. the implementation of a data crawler, data transformer, dataset for storing FDB data, construction of a P&D IE keyword template, novel forward analyser, novel backward analyser and STS approach).

A prototype system architecture overview was presented in Chapter 5. The six sections (Section 5.3 to Section 5.8) in the same chapter described each of the components in the architecture that implements the 7-phase methodology. Lastly, the Graphical User Interface (GUI) of the data crawler, novel forward analyser and novel backward analyser was presented.

Chapter 6 conducted a series of experiments for the novel forward analysis and backward analysis introduced in this research – two experiments for the forward analysis and another two for the backward analysis. The outcomes of the experiments were also discussed and it appeared to be supporting all the experiments' hypotheses. Chapter 7

also presented experiments using the Semantic Textual Similarity (STS) approach to test whether the STS approach can be adopted in the detection of potentially illegal FDB activities. The outcomes of the experiments were also discussed and they appeared to be supporting the experiments' hypotheses. Finally, Chapter 8 concluded the thesis, highlighting the contributions and discussing the potential future directions of this research.

8.3 Contributions Summary

The key contribution of this research is the creation of a semi-automated template-based IE prototype system, namely Financial Discussion Boards Miner (FDBM), that can be used for the detection of potentially illegal P&D or other financial crimes such as Insider Information (II) on share price based FDBs.

The contributions are elaborated as follows:

1. A semi-automated stock market surveillance prototype system, namely FDBs Miner (FDBM), developed based on the novel methodologies devised in Contribution 5.
2. A crawler component in the FDBM prototype system that is capable of crawling semi-structured data from FDBs.
3. A data transformer component that can pre-process and transform the FDBs related semi-structured data, collected in Contribution 1, into structured data.
4. A novel FDB dataset (FDB-DS) that contains FDB artefact data such as ticker symbols, FDB comments, share prices, broker ratings and director deals data that belonged to all the 941 companies (picked from the FTSE-100 and FTSE AIM All-Share indices) on the London Stock Exchange (LSE, 2017).
5. Two novel methodologies, namely forward analysis and backward analysis, that utilise IE and Moving Average (MA) techniques have been established for the detection of potentially illegal P&D activities on FDBs involving FDB comments and share prices.

6. A predefined IE keyword template constructed based on P&D financial crime. The template contains keywords, phrases and sentences, that are commonly used by fraudsters on FDBs. It can be employed in real-world scenarios by the relevant authorities.
7. A system that is capable of performing anomaly detection in terms of potential P&D activities on FDBs and other potential financial crimes on FDBs provided if the IE keywords templates are devised.
8. A system that can integrate with a Semantic Textual Similarity (STS) approach to compare the semantic similarity between the FDB comments and the keywords, phrases and short sentences in the P&D IE keyword template. This aids the relevant authorities to investigate potentially illegal FDB comments based on different semantic similarity thresholds in addition to the two novel methodologies described in Contribution 5.

8.4 Future Directions

The FDBM prototype system in this research does not represent the final solution or definitive version of the stock market surveillance system in relation to share price based FDBs. There are several potential future works that can be undertaken in order to enhance the functionality and extend the features of FDBM for surveillance of potentially illegal activities on FDBs.

Possible functionality and feature related future works are listed below:

- The creation of other potential financial crimes in relation to the share price based FDBs such as Insider Information (II). II is not as popular as P&D in relation to share price based FDBs. However, it is also possible that it might occur on FDBs. Hence, if a predefined IE keyword template can be devised and validated by experts in the field, FDBM can also be used to detect potentially illegal II FDB comments.

- The director deals and broker ratings FDB artefact data were collected in the process of data crawling in Section 5.3. These data may possibly be involved in suspicious FDB comment flagging alongside comments and share prices.
- At the time the artefacts data were collected, FDB comments' RSS feed (in the form of XML files – semi-structured data) were crawled and transformed into structured data. However, London South East FDB (LSE-FDB, 2017) has since removed the RSS feed option on the website. Hence, as part of future work, the data crawler component could be modified in order to crawl HTML files to obtain the FDB comments instead. This would also be suitable for crawling the comments on ADVFN (ADVFN, 2017) because it has never provided any RSS feed.
- FDB comments from other countries such as HotCopper Australia (HotCopper, 2017) may also be suitable for investigations through the use of the FDBM prototype system, as long as it is capable of crawling data in the supported format such as RSS feed or the HTML format as part of the future work.
- FDB-FS may be published and made public for further research by researchers in the field.
- The FDBM prototype system can be revised and improved by constructing a richer Graphical User Interface (GUI) for better visualisation such as visualising the price hike thresholds in the forward analysis, moving average thresholds in the backward analysis and semantic similarity thresholds in the Semantic Textual Similarity (STS) approach.
- It is also possible to create a more sophisticated set of filters on the backward analyser GUI to allow advanced filtration of data for investigation purposes.

8.5 Overall Conclusion

In conclusion, this thesis reviewed the stock exchange in the UK, namely the London Stock Exchange (LSE, 2017); three share price based FDBs in the UK, namely ADVFN (ADVFN, 2017), London South East (LSE-FDB, 2017) and Interactive Investors (III, 2017); functions of the stock market regulatory agencies in both the US and UK, namely The

Security and Exchange Commission (SEC) and the Financial Conduct Authority (FCA) respectively; and, Pump and Dump (P&D) financial crimes and existing studies in relation to FDBs. The three types of data structures that can be collected from the Internet and the challenges of collecting and processing them have also been discussed. Two Information Extraction (IE) approaches, namely a Knowledge Engineering approach and an Automatic Learning approach, have also been reviewed.

A FDBM prototype system was presented. It implemented a methodology (which comprises a novel forward analysis and backward analysis) and architecture (which comprises seven components). The FDBM prototype system can aid the relevant authorities to proactively monitor and detect potentially illegal activities on the FDBs. With the creation of the FDBM prototype system, the problems faced by relevant authorities and FDB moderators of not having enough manpower or time to monitor share price based FDBs for potentially illegal activities could be solved; this would then directly benefit novice investors, who are normally the ones deceived by financial crime fraudsters.

References

Ackert L. F., Jiang, L., Lee H. S. and Liu, J. (2016) 'Influential investors in online stock forums.' *International Review of Financial Analysis 45,* pp. 39–46.

ADVFN PLC, (2017) *ADVFN*. [Online] [Accessed on 6th September 2017] http://uk.advfn.com

Akismet (2017) *Akismet*. [Online] [Accessed on 6th September 2017] https://akismet.com

Alić, I. (2015) 'Supporting Financial Market Surveillance: An IT Artifact Evaluation' *BLED 2015 Proceedings*, Paper 36.

Aljameel, S. S., O'Shea, J. D., Crockett, K. A., Latham, A. and Kaleem, M. (2017) 'Development of An Arabic Conversational Intelligent Tutoring System for Education of Children with ASD'

Antweiler, W. and Frank, M. Z. (2004) 'Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards.' *The Journal of Finance*, 59(3), pp. 1259–1294.

Appelt, E. and Israel, D. (1999) 'Introduction to Information Extraction.'

Australian Stock Exchange. (2017) 'Australian Stock Exchange' [Online] [Accessed on 6th September 2017] http://www.asx.com.au

Bank of England. (2017) 'Prudential Regulations' [Online] [Accessed on 6th September 2017] https://www.bankofengland.co.uk/prudential-regulation

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,* pp. 50-64

Bettman, J., Hallett, A. and Sault, S. (2011) 'Rumortrage: Can Investors Profit on Takeover Rumors on Internet Stock Message Boards?'

Bird, S. (2006) 'Nltk: the natural language toolkit. *In* Proceedings of the COLING/ACL on Interactive presentation sessions, COLING-ACL '06, *Association for Computational Linguistics*, pp. 69–72, Stroudsburg, PA, USA.

Bouraoui, T. (2015) 'Does 'Pump and Dump' Affect Stock Markets?' *International Journal of Trade, Economics and Finance*, 6(1).

Campbell, J. A. (2001) 'In and Out Scream and Shout: An Internet Conversation about Stock Price Manipulation.' *Proceedings of the 34th Hawaii International Conference on System Sciences*, pp. 1–10.

Campbell, J.A. and Cecez-Kecmanovic, D. (2011) 'Communicative practices in an online financial forum during abnormal stock market behavior.' *Information and Management*, 48, pp. 37-52.

Chiticariu, L., Li, Y. and Reiss, R. F. (2013) '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*, pp. 827–832, Seattle, Washington.

Cook, D. O. and Lu, X. (2009) 'Noise, Information, and Rumors: Internet Boards Messages Affect Stock Returns.' University of Alabama.

Cowie, J. and Lehnert, W. (1996) 'Information Extraction.' *Communications of the ACM*, 39(1), pp. 80–91.

Cunningham H. (2006) 'Information Extraction, Automatic.' *In* Keith Brown, *(Editor-in-Chief) Encyclopedia of Language & Linguistics*, Second Edition, 5, pp. 665-677.

Cybenko, G., Giani, A. and Thompson, P. (2002) 'Cognitive Hacking: A Battle for the Mind.'

Delort, J. Y., Arunasalam, B. and Leung, H. (2011) 'The Impact of Manipulation in Internet Stock Message Boards.' *International Journal of Banking and Finance*, 8(4), pp. 1–18.

Delort, J. Y., Arunasalam, B. and Paris, C. (2011) 'Automatic Moderation of Online Discussion Sites.' *International Journal of Electronic Commerce*, 15(3), pp. 9–30.

Dzikevičius, A. and Šaranda, S. (2010) 'EMA versus SMA Usage to Forecast Stock Markets: The Case of S&P 500 and OMX Baltic Benchmark.' *Business: Theory and Practice*, 11(3), pp. 248-255.

Fellbaum, C. (2006) 'WordNet(s).' *In* Keith Brown, *(Editor-in-Chief) Encyclopedia of Language & Linguistics*, Second Edition, 13, pp. 665-670. Oxford: Elsevier.

Felton, J. and Kim, J. (2002) 'Warnings from the Enron Message Board.' *Journal of Investing*, 11(3), pp. 29-52.

Financial Conduct Authority. (2016) *MAR 1.8 Market abuse (dissemination).* [Online] [Accessed on 9th March 2017] https://www.handbook.fca.org.uk/handbook/MAR/1/8.html?date=2016-06-01

Financial Conduct Authority. (2016) *Number of STORSs received.* [Online] [Accessed on 9th March 2017] https://www.fca.org.uk/markets/market-abuse/suspicious-transaction-and-order-reports/number-stors-received

Financial Conduct Authority. (2016) *Number of Suspicious Transaction Reports (STRs) Received under the Market Abuse Directive (Directive 2003/6/EC) by the FCA 2007 - 2nd July 2016*. [Online] [Accessed on 9th March 2017] https://www.fca.org.uk/publication/data/number-of-suspicious-transaction-reports.pdf

Financial Conduct Authority. (2017) 'Financial Conduct Authority' [Online] [Accessed on 6th September 2017] https://www.fca.org.uk

Financial Service Authority. (2017) 'Financial Service Authority' [Online] [Accessed on 6th September 2017] http://www.fsa.gov.uk

Financial Times. (2017) *Master list of boiler room victims found.* [Online] [Accessed on 9th March 2017] https://www.ft.com/content/518efafa-01f1-11e0-b66c-00144feabdc0

Foltz, P., Kintsch, W. and Landauer, T. (1998) 'The measurement of textual coherence with latent semantic analysis.' *Disc. Proc. 25*, 2–3, pp285–307.

Gomaa W. H. and Fahmy A. A. (2013) 'A Survey of Text Similarity Approaches.' *International Journal of Computer Applications (0975 – 8887)*, 68(13).

Halifax UK. (2017) *Regulatory Information.* [Online] [Accessed on 23rd March 2017] https://www.halifax.co.uk/sharedealing/important-information /regulatory-information/

HotCopper (2017) *HotCopper*. [Online] [Accessed on 9th September 2017] https://hotcopper.com.au

HotStocked.com. (2016) *How to Become a Better Trader in 7 Days or Less!* [Online] [Accessed on 1st March 2017] http://newsletter.hotstocked.com/newsletters

IBM. (2017) *Analyzing big data with Text Analytics*. [Online] [Accessed on 1st December 2017] https://www.ibm.com/support/knowledgecenter /en/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.text.doc/doc/ana_txtan_intro.html

Interactive Investor Plc. (2017) *Interactive Investor*. [Online] [Accessed on 6th September 2017] http://www.iii.co.uk

Investopedia. (2017) *Boiler Room.* [Online] [Accessed on 8th March 2017] https://www.investopedia.com/terms/b/boilerroom.asp

Knott, E. and Owda, M. (2012) 'The detection of potentially illegal activity on financial discussion boards using information extraction.' *2nd International Conference on Cybercrime, Security and Digital Forensics*, London.

Landauer, T. and Dumais, S. (1997) 'A solution to platos problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge.' *Psych. Rev*. 104(2), pp211– 240.

Landauer, T., Foltz, P. and Laham, D. (1998) 'Introduction to latent semantic analysis.' *Dis. Proc*. *25*, 2(3), pp. 259–284.

Lee, M. C., Chang, J. W. and Hsieh, T. C. (2014) 'A Grammar-Based Semantic Similarity Algorithm for Natural Language Sentences.' *The Scientific World Journal*.

Lee, P. S., Owda, M. and Crockett, K. (2018) '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.

Leinweber, D. J. and Madhavan, A. N. (2001) 'Three Hundred Years of Stock Market Manipulations.' *Journal of Investing* pp. 7–16.

Leung, H., and Ton, T. (2015) 'The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks.' *Journal of Banking & Finance*, pp. 37–55.

Lewis, M. (2001) *Jonathan Lebed's Extracurricular Activities. The New York Times*. [Online] [Accessed on 6th September 2017] http://www.nytimes.com /2001/02/25/magazine/jonathan-lebed-s-extracurricular-activities.html? pagewanted=all&src=pm

Limanto et al. (2005) 'An Information Extraction Engine for Web Discussion Forums.' *Nanyang Technological University*, Singapore. ACM 1-59593-051-5/05/0005

London South East Limited. (2017) *London South East*. [Online] [Accessed on 6th September 2017] http://www.lse.co.uk

London Stock Exchange. (2017) *London Stock Exchange*. [Online] [Accessed on 6th September 2017] http://www.londonstockexchange.com

Masterson, D. and Kushmerick, N. (2003) 'Information Extraction from Multi-Document Threads.'

Microsoft. (2017) *FAST Search Server 2010 for SharePoint.* [Online] [Accessed on 1st December 2017] https://technet.microsoft.com/en-us/library/ ee781286(v=office.14).aspx

Miller. G. A. (1995) 'WordNet: a lexical database for English' *Communications of the ACM*, 38(11):41.

NASDAQ Stock Exchange. (2017) 'Australian Stock Exchange' [Online] [Accessed on 6th September 2017] http://www.nasdaq.com

New York Stock Exchange. (2017) 'New York Stock Exchange' [Online] [Accessed on 6th September 2017] https://www.nyse.com/index

O'Shea, J, Bandar, Z, and Crockett, K. (2014) 'A new benchmark dataset with production methodology for short text semantic similarity algorithms.' *ACM Transactions on Speech and Language Processing (TSLP)* 10.4:19

Owda, M., Lee, P. S. and Crockett, K. (2017) 'Financial Discussion Boards Irregularities Detection System (FDBs-IDS) using Information Extraction.' *Intelligent Systems Conference 2017*, London.

Phillippsohn, S. (2001) 'Trends In Cybercrime — An Overview Of Current Financial Crimes On The Internet.' *Computers & Security*, 20, pp. 53-69.

Problogger. (2017) *What is RSS?* [Online] [Accessed on 10th June 2017] http://www.problogger.net/what-is-rss

Raiyn, J. and Toledo, T. (2014) 'Real-Time Road Traffic Anomaly Detection.' *Journal of Transportation Technologies*, 4(3), pp. 256-266.

Raiyn, J. and Toledo, T. (2014) 'Real-time Short-term Forecasting Based on Information Management.' *Journal of Transportation Technologies*, 4, pp. 11-21.

Renault, T. (2014) 'Pump-and-dump or news? Stock market manipulation on social media.'

Renault, T. (2017) 'Market Manipulation and Suspicious Stock Recommendations on Social Media.'

Riem, A. (2001) 'Cybercrimes of the 21st Century: Crimes against the individual — Part 1.' *Computer Fraud & Security*," 6, pp. 13–17.

Ross, S. (2017) *What is the difference between Exponential Moving Average (EMA) and Weighted Moving Average?* [Online] [Accessed on 21st August 2017]

https://www.investopedia.com/ask/answers/122214/what-difference-between-exponential-moving-average-ema-and-weighted-moving-average.asp

Rzepka, R., Amaya, Y., Yatsu, M. and Araki, K. (2015) 'Automatic Narrative Humor Recognition Method Using Machine Learning and Semantic Similarity Based Punchline Detection.' *International Workshop on Chance Discovery, Data Synthesis and Data Market*.

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

Sandbank, T., Shmueli-Scheuer, M., Konopnicki, D., Herzig, J., Richards, J. and Piorkowski, D. (2017) 'Detecting Egregious Conversations between Customers and Virtual Agents'

Seo, K., Choi, J. and Choi, Y. (2009) 'Research about Extracting and Analyzing Accounting Data of Company to Detect Financial Fraud.' *Intelligence and Security Informatics*, pp. 200–202.

Siering, M. (2013) 'All Pump, No Dump? The Impact of Internet Deception on Stock Markets.' *ECIS 2013 Completed Research*, 115.

Soderland, S. (1999) 'Learning Information Extraction Rules for Semi-structured and Free Text.'

Software Garden, Inc. (2004) *What is RSS?* [Online] [Accessed on 5th April 2017] http://rss.softwaregarden.com/aboutrss.html

StockCharts (2017) *Moving Averages – Simple and Exponential*. [Online] [Accessed on 12th December 2017] http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators:moving_averages

Symantec. (2017) *Symantec.cloud effectiveness user's guide.* [Online] [Accessed on 4th April 2017] https://support.symantec.com/ en_US/article.TECH222392.html

U.S. Securities and Exchange Commission (2015) *SEC Files Charges in Multi-Million Dollar Market Manipulation*. [Online] [Accessed on 18th July 2017] https://www.sec.gov/news/pressrelease/2015-279.html

U.S. Securities and Exchange Commission (2017) *What We Do*. [Online] [Accessed on 25th August 2017] https://www.sec.gov/Article/whatwedo.html

U.S. Securities and Exchange Commission. (2017) *Financial Navigating in the Current Economy: Ten Things to Consider Before You Make Investing Decisions.* [Online] [Accessed on 1st March 2017] https://www.sec.gov/investor/pubs/tenthingstoconsider.htm

U.S. Securities and Exchange Commission. (2017) *SEC Announces Charges in Massive Telemarketing Boiler Room Scheme Targeting Seniors.* [Online] [Accessed on 9th March 2017] https://www.sec.gov/news/press-release/2017-124

US Security and Exchange Commission (2009) *SEC Charges Eight Participants in Penny Stock Manipulation Ring*. [Online] [Accessed on 8th August 2017] http://www.sec.gov/litigation/litreleases/2009/lr21053.htm

vBulletin Solutions Inc. (2017) *vBulletin Community Forum help.* [Online] [Accessed on 26th March 2017] https://www.vbulletin.com/forum/help?faq=vb3_board_usage#faq_vb3_forums_threads_posts

Visual Capitalist (2017) *The 20 Largest Stock Exchanges in the World*. [Online] [Accessed on 25th August 2017] http://www.visualcapitalist.com/20-largest-stock-exchanges-world

Wired. (2017) Google *Says Its AI Catches 99.9 Percent of Gmail Spam.* [Online] [Accessed on 20th March 2017] https://www.wired.com/2015/07/google-says-ai-catches-99-9-percent-gmail-spam/

Wolfram, M. S. A. (2010) 'Modelling the Stock Market using Twitter.'

World Wide Web Foundation. (2017) *History of the Web.* [Online] [Accessed on 5th April 2017] https://webfoundation.org/about/vision/history-of-the-web/

Appendix A

Lee, P. S., Owda, M. and Crockett, K. (2014) 'A Financial Crime Analysis Methodology for Financial Discussion Boards using Information Extraction Techniques.' *7th Manchester Metropolitan University Postgraduate Research Conference*, Manchester Metropolitan University, Manchester.

A FINANCIAL CRIME ANALYSIS METHODOLOGY FOR FINANCIAL DISCUSSION BOARDS USING INFORMATION EXTRACTION TECHNIQUES (Oral Presentation)

Author: Pei-Shyuan Lee, Dr Majdi Owda, Dr Keeley Crockett
Institution: Manchester Metropolitan University, United Kingdom
Email: pei-shyuan.lee@mmu.ac.uk, m.owda@mmu.ac.uk, k.crockett@mmu.ac.uk

Keywords: Financial discussion boards, pump and dump, insider information, information extraction, artificial intelligence

Abstract

Financial Discussion Boards (FDBs) are places where investors involved in discussions for financial knowledge exchange. There are various popular FDBs, which specifically allow the discussions of share prices. Financial crimes such as Pump and Dump and Insider Information can be found in these FDBs [1,2,3]. 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.

There were a number of attempts to classify labelled FDB comments using Artificial Intelligence (AI) techniques to statistically analyse the impact towards stock market. Potentially illegal FDB comments were found manipulative and positively related to the market returns, volatility and trading volumes [4,5,6]. However, no existing research look into comments and share prices simultaneously to avoid false signs of illegal comments during the detection of potentially illegal comments. There was also very little attempts to automate financial surveillance systems for detecting potentially illegal comments on FDBs to solve manual moderations that require enormous efforts.

This work proposes a novel methodology for the creation of a template based prototype system which uses Information Extraction (IE) techniques to first extract information from an unstructured or semi-structured data source (such as ticker symbols, FDB comments and share prices) into a structured data format and then flag potentially illegal comments by looking at the share prices simultaneously. This work also attempts to automate the prototype system and implement AI techniques for comments classification.

References

[1] Campbell, J.A. (2001). In and Out Scream and Shout: An Internet Conversation about Stock Price Manipulation. *Proceedings of the 34th Hawaii International Conference on System Sciences* (pp. 1–10).
[2] Delort, J. Y., Arunasalam, B., & Leung, H. (2011). The Impact of Manipulation in Internet Stock Message Boards. *International Journal of Banking and Finance*, *8*(4), 1–18.
[3] Knott, E., & Owda, M. (2012). The detection of potentially illegal activity on financial discussion boards using information extraction. *2nd International Conference on Cybercrime, Security and Digital Forensics, London, UK.*
[4] Cook, D. O., & Lu, X. (2009). Noise, Information, and Rumors: Internet Boards Messages Affect Stock Returns. University of Alabama.
[5] Sabherwal, S., Sarkar, S. K., & Zhang, Y. (2011). Do Internet Stock Message Boards Influence Trading? Evidence from Heavily Discussed Stocks with No Fundamental News. *Journal of Business Finance & Accounting*, *38*(9-10), 1209–1237.
[6] Bettman, J., Hallett, A., & Sault, S. (2011). Rumortrage: Can Investors Profit on Takeover Rumors on Internet Stock Message Boards?

A Financial Crime Analysis Methodology for Financial Discussion Boards using Information Extraction Techniques

Pei-Shyuan Lee, Dr. Majdi Owda, and Dr. Keeley Crockett

Faculty of Science and Engineering, Manchester Metropolitan University

pei-shyuan.lee@stu.mmu.ac.uk, m.owda@mmu.ac.uk, k.crockett@mmu.ac.uk

Introduction

Financial Discussion Boards (FDBs) are places where investors involved in discussions for financial knowledge exchange. Financial crimes such as Pump and Dump and Insider Information can be found in FDBs [1]. Comments like "This is the right time let's start pumping this share" can reveal a hidden potential illegal activity of Pump and Dump. For example, in 2009, SEC charged eight participants due to their involvement in penny stock manipulation during the year 2006 to 2007 [2]. Most of them met each other through the popular penny stock message board (i.e. InvestorsHub.com) and carried out the Pump and Dump scheme. This shows criminals are using public forums to organise crimes.

Delort et al [3] found that share price pumping on HotCopper, Australian's largest internet stock message board, was positively related to market returns, volatility and volume. Delort et al [3] concluded that manual supervision of forum posting activities by forum moderators are not sufficient to protect users against manipulation. FDBs need to be monitored and such financial crimes need to be prevented not only by the forum moderators and relevant law enforcement agencies but also with the help of crime detection software.

Financial Discussion Board Miner

This research aims to design a novel methodology for the development of a system called Financial Discussion Board Miner (FDBM) which uses Information Extraction (IE) [4] techniques to first extract information from an unstructured or semi-structured data source (such as FDB comments and share prices) into a structured data format and then locate potentially illegal comments. IE systems are knowledge-intensive [5] since 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 user [6]. The proposed templates will display a summary of information from a number of inter-linked sources (i.e. FDB comments and share prices) allowing filtering of potentially illegal comments to take place.

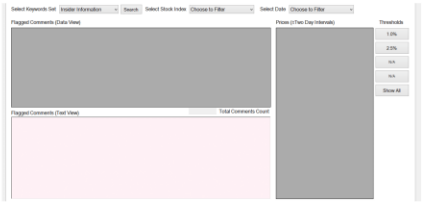

Figure 1. Prototype system

FDBM will search and identify potentially illegal comments by looking at comments and share price simultaneously to avoid false signs of illegal comments. One novelty of the FDBM is the ability to analyse both comments and share prices together. FDBM will capture all available textual comments automatically from popular share price based FDBs [8,9] in the United Kingdom and extract the live share prices from ADVFN [10] using web crawler. The dataset will be captured over 12 weeks and stored in a database for analysis.

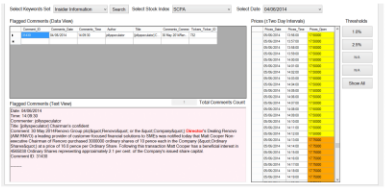

Figure 2. Prototype system—Filtering potentially illegal comments

The semantically understandable artefacts (i.e. the artefacts that can be processed by computers) play an important role in the formulation of IE templates, since they are unique artefacts such as stock ticker names, usernames, date, time, price and user sentiments. IE templates will be constructed using financial crimes' related keywords that commonly used in FDBs in addition to the semantically understandable artefacts. Artificial Intelligence (AI) techniques such as Naïve Bayes and bag-of-words model will initially be used for "good" and "bad" comments classification. "Bad" comments could then be classified further into specific financial crimes categories such as Pump and Dump and Insider Information using the predefined IE templates.

Conclusion

The proposed prototype system, FDBM, can be helpful to flag and prevent potential crimes on share price based FDBs using IE techniques. Taking both text comments and share prices into account can also resolve the problem of having false signs of illegal text comments by looking at the share prices at the time of posting that illegal comment. A methodology for extracting data in real time from the three sources [8,9,10] has been developed along with the FDBM framework. Currently, data is being extracted through the system to build the dataset.

References

[1] Campbell, J.A. (2001). In and Out Scream and Shout: An Internet Conversation about Stock Price Manipulation. Proceedings of 34th Hawaii International Conference on System Sciences (pp. 1–10).
[2] US Security & Exchange Commission (2001). Pump and Dump Schemes [Online]. Available from: www.sec.gov/answers/pumpdump.htm [17 Nov 2012]
[3] Delort, J. Y., Arunasalam, B., & Leung, H. (2011). The Impact of Manipulation in Internet Stock Message Boards. International Journal of Banking and Finance, 8(4), 1–18.
[4] Cowie, J., & Lehnert, W. (1996). Information Extraction. Communications of the ACM, 39(1), 80–91.
[5] Soderland, S. (1999). Learning Information Extraction Rules for Semi-structured and Free Text.
[6] Cunningham H (2006), Information Extraction, Automatic. In: Keith Brown, (Editor-in-Chief) Encyclopedia of Language & Linguistics, Second Edition, volume 5, pp. 665-677. Oxford: Elsevier.
[7] Limanto H.Y. et al. (2005). An Information Extraction Engine for Web Discussion Forums. Nanyang Technological University, Singapore. ACM 1-59593-051-5/05/0005.
[8] London South East Limited (2012). London South East [Online]. Available: http://www.lse.co.uk [6 Sept 2013]
[9] Interactive Investor Plc. (2012). Interactive Investor [Online]. Available: http://www.iii.co.uk [6 Sept 2013]
[10] ADVFN PLC (2013). ADVFN [Online]. Available: http://uk.advfn.com [6 Sept 2013]

Appendix B

Owda, M., Lee, P. S. and Crockett, K. (2017) 'Financial Discussion Boards Irregularities Detection System (FDBs-IDS) using Information Extraction.' *Intelligent Systems Conference 2017*. London, 7th-8th September 2017.

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.

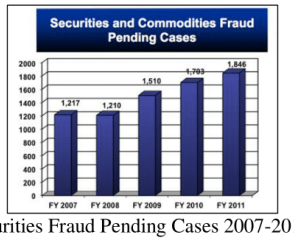

Fig. 1. Securities Fraud Pending Cases 2007-2011 [16]

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.

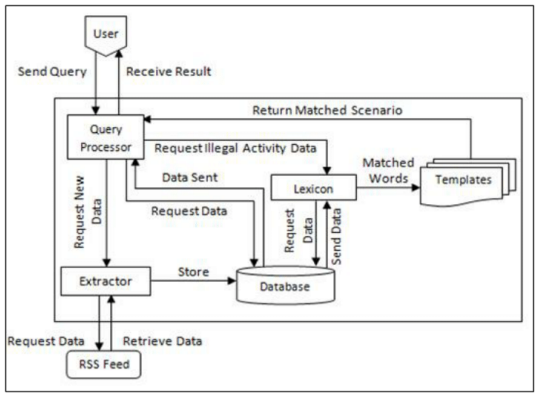

Fig. 2. Component Diagram for the Proposed FDBs-IDS

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.

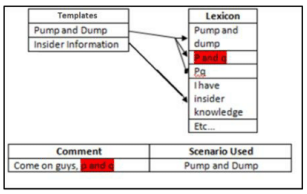

Fig. 3. Example Simple Template Structure

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.

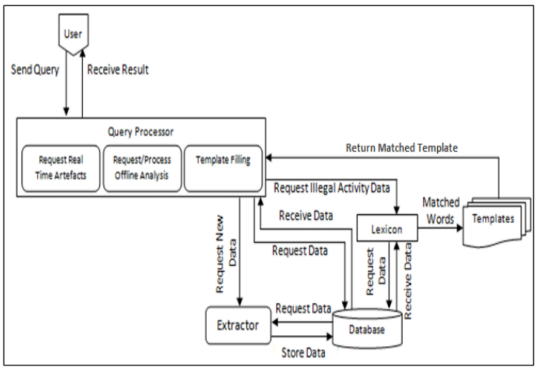

Fig. 4. Query Processor

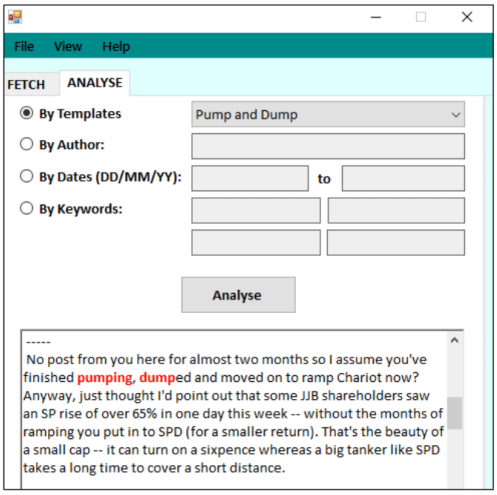

Fig. 5. Screenshot of FDBs-IDS System

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.

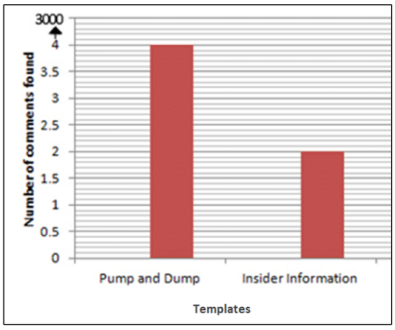

Fig. 6. Results for two templates

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.

Appendix C

Lee, P. S., Owda, M., & Crockett, K. (2018) 'The detection of fraud activities on the stock market through forward analysis methodology of financial discussion boards.' *Future of Information and Communications Conference*. Singapore, 5th-6th April 2018.

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. 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

***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.

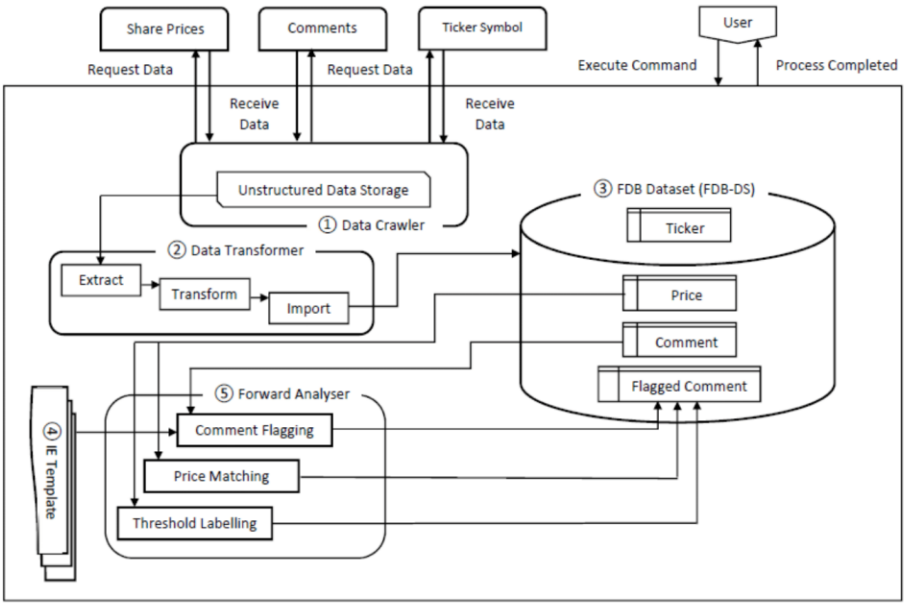

Fig. 1. Architecture Overview Diagram.

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%

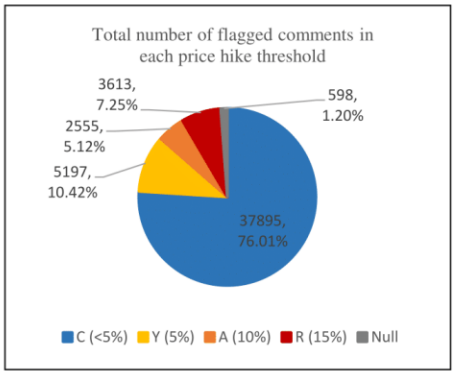

Fig. 2. Total number and percentage of each threshold.

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.

Appendix D

Lee, P. S., Owda, M. and Crockett, K. (2018) 'Methodologies for resolving false positives during the detection of fraudulent activities on the stock market through forward and backward analysis of financial discussion boards.' *International Journal of Advanced Computer Science and Applications*, 9(1).

Methodologies for resolving false positives during the detection of fraudulent activities on the stock market through forward and backward analysis 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. 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 [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 are 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). 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 [5, 6, 7, 8, 9, 10, 11, 12]. Research from recent years highlighted that the comments on FDBs were found manipulative and positively related to the market returns, volatility and trading volumes [13, 14, 15, 16, 17]. However, there is very little attempt [11, 12] 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 [18], 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 [19]. Therefore, IE techniques are used in this research to extract and analyse these data. IE has been used in other areas such as accounting [20] and search engine [21]. However, other than the initial work described in [12] and [22], 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 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 [7]. 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) [7, 8, 9].
- 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 [8].
- In 2009, eight participants were charged by Security Exchange Commission (SEC) for being involved in penny stock manipulation [10]. These wrongdoers met

each other through InvestorsHub (now owned by ADVFN [3]), a popular penny stock message board. Followed by carrying out a P&D scheme throughout the year of 2006 and 2007.

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 [23] as the process of extracting information automatically into a structured data format from an unstructured or semi-structured data sources. It was suggested [24] that there is a need for systems that extract information automatically from text data. IE is not Information Retrieval (IR) [25]. 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) [26]. 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 [26]. This paper focuses on IE implementation since it is designed to support the financial market surveillance authorities.

Text mining was described [27] 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 [28], 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. A text mining based study was conducted [29] on the Twitter found it to be able to predict stock prices. Stock price trends were also being successfully forecasted via press releases using text mining techniques [30].

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 symbol.

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 [31], 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. Some research [13, 14, 15, 16, 17] have found the relation between FDB comments and market performance. FDB comments can be manipulative and affect the share prices.

In [11], 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 [32]. 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 [12] 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 process 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.

The authors in [17] 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 [33], 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.

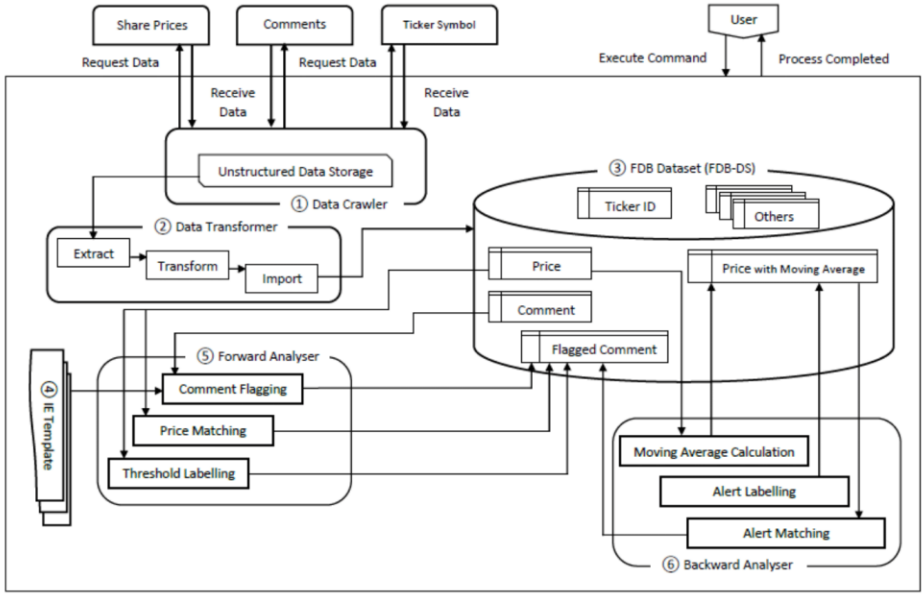

Fig. 1. Architecture Overview Diagram.

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 [1], III [2] and ADVFN [3]) 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) [1] 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.

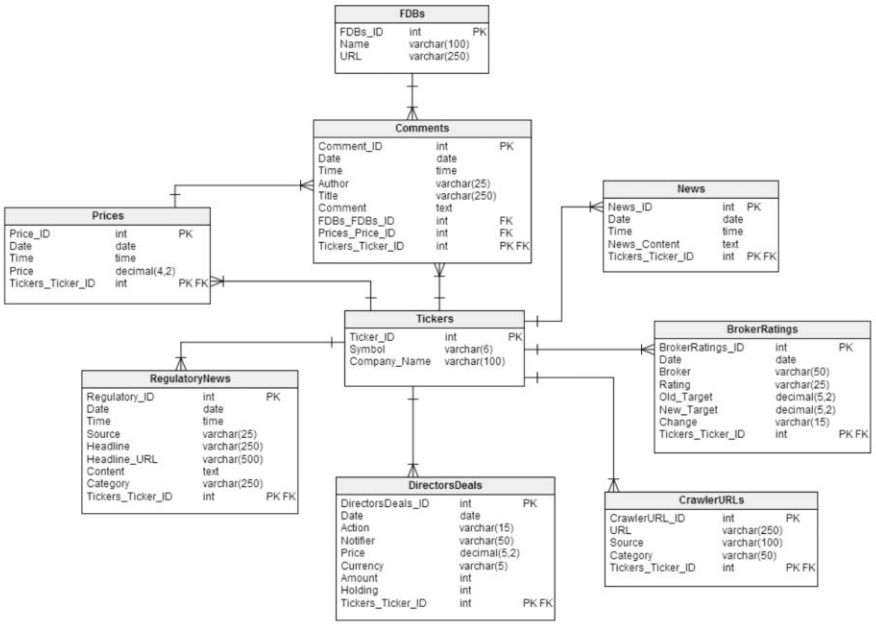

Fig. 2. FDB Dataset Structure.

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 be easily modified whenever needed. The IE keyword template consists of a series of keywords and phrases that were thoroughly researched [6, 34, 35, 36] 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 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 [1] and III [2]. 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, 34, 35, 36]. 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%

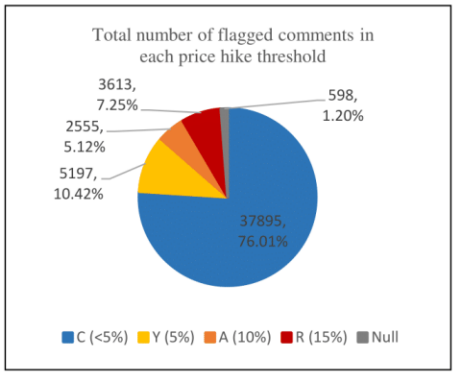

Fig. 3. Total number and percentage of each threshold.

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 [37]. 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:

$$SMA = \frac{p_1 + p_2 + \cdots + p_n}{n} \quad [38]$$

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

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

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

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

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

[6] 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.

[7] 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, November 2017.

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

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

[10] 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, November 2017

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

[12] 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.

[13] 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.

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

[15] 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.

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

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

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

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

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

[26] 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.

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

[28] Gottlieb, J., "Social Media Users Actively Seek and Share Advice," Available at: http://www.roiresearch.com/blog/post/2011/01/04/Social-Media-Users-Actively-Seek-and-Share-Advice.aspx, November 2017

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

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

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

[32] HotCopper, "HotCopper" [Online]. Available: https://hotcopper.com.au, November 2017

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

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

[35] 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.

[36] 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.

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

[38] StockCharts, "Moving Averages – Simple and Exponential" [Online] Available: http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators:moving_averages, November 2017

Appendix E

A screenshot of the email reply from the owner of LSE-FDB.

phil.thomas@lse.co.uk <phil.thomas@lse.co.uk> 24 September 2013 at 17:01
To: sshyuann@gmail.com

Hi Shyuan,

Thank you for your e-mail. Usually, we say no to this sort of thing - but it does sound interesting. I would prefer to know a bit more about your methodology before agreeing; for instance, were you planning to crawl one share at a time, and if so, how would you decide which shares to crawl?

As I mentioned, the project sounds interesting - so, assuming the methodology will not cause us any undue problems, I would agree on the basis that we received a copy of your findings. To counter that, I would try to provide assistance where possible.

I'm not especially happy about the idea of the website pages being crawled directly, but you may not be aware that we run an RSS service for the chat. Examples include:

http://www.lse.co.uk/chat/rss/

http://www.lse.co.uk/chat/RBS/

I would feel a lot more comfortable if you used this instead of crawling directly. I can arrange for the share price at the time of posting to appear in the feed also, as I'm aware it's not there are present.

Let me know your thoughts, or if you have any questions.

Kind Regards,

Phil Thomas

Appendix F

threshold_Yes_No * sma3_threshold_YesNo Crosstabulation

		sma3_threshold_YesNo		
		0	1	Total
threshold_Yes_No	0	36807	1686	38493
	1	9878	1487	11365
Total		46685	3173	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1115.521[a]	1	.000		
Continuity Correction[b]	1114.061	1	.000		
Likelihood Ratio	955.839	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	1115.498	1	.000		
N of Valid Cases	49858				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 723.28.

b. Computed only for a 2x2 table

Symmetric Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval		Pearson's R	.150	.005	33.779	.000[c]
Ordinal by Ordinal		Spearman Correlation	.150	.005	33.779	.000[c]
N of Valid Cases			49858			

threshold_Yes_No * sma5_threshold_YesNo Crosstabulation

		sma5_threshold_YesNo		
		0	1	Total
threshold_Yes_No	0	36432	2061	38493
	1	9823	1542	11365
Total		46255	3603	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	882.971[a]	1	.000		
Continuity Correction[b]	881.746	1	.000		
Likelihood Ratio	771.575	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	882.953	1	.000		
N of Valid Cases	49858				

Symmetric Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	by	Pearson's R	.133	.005	29.981	.000[c]
Ordinal by Ordinal	by	Spearman Correlation	.133	.005	29.981	.000[c]
N of Valid Cases			49858			

threshold_Yes_No * wma1_threshold_YesNo Crosstabulation

		wma1_threshold_YesNo		
		0	1	Total
threshold_Yes_No	0	37770	723	38493
	1	10552	813	11365
Total		48322	1536	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	817.788[a]	1	.000		
Continuity Correction[b]	816.022	1	.000		
Likelihood Ratio	679.730	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	817.771	1	.000		
N of Valid Cases	49858				

Symmetric Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval	by	Pearson's R	.128	.006	28.834	.000[c]
Interval						
Ordinal	by	Spearman Correlation	.128	.006	28.834	.000[c]
Ordinal						
N of Valid Cases			49858			

threshold_Yes_No * wma3_threshold_YesNo Crosstabulation

		wma3_threshold_YesNo		
		0	1	Total
threshold_Yes_No	0	37332	1161	38493
	1	10206	1159	11365
Total		47538	2320	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1020.068[a]	1	.000		
Continuity Correction[b]	1018.450	1	.000		
Likelihood Ratio	860.131	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	1020.047	1	.000		
N of Valid Cases	49858				

Symmetric Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	by	Pearson's R	.143	.005	32.270	.000[c]
Ordinal by Ordinal	by	Spearman Correlation	.143	.005	32.270	.000[c]
N of Valid Cases			49858			

threshold_Yes_No * sma5_threshold_YesNo Crosstabulation

Count

		sma5_threshold_YesNo		
		0	1	Total
threshold_Yes_No	0	36941	1552	38493
	1	9999	1366	11365
Total		46940	2918	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1015.955[a]	1	.000		
Continuity Correction[b]	1014.506	1	.000		
Likelihood Ratio	869.726	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	1015.935	1	.000		
N of Valid Cases	49858				

Symmetric Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval	by	Pearson's R	.143	.005	32.203	.000[c]
Interval						
Ordinal	by	Spearman Correlation	.143	.005	32.203	.000[c]
Ordinal						
N of Valid Cases			49858			

threshold_Yes_No * ema1_threshold_YesNo Crosstabulation

		ema1_threshold_YesNo		
		0	1	Total
threshold_Yes_No	0	36925	1568	38493
	1	10324	1041	11365
Total		47249	2609	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	457.733[a]	1	.000		
Continuity Correction[b]	456.708	1	.000		
Likelihood Ratio	404.356	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	457.723	1	.000		
N of Valid Cases	49858				

Symmetric Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval	by	Pearson's R	.096	.005	21.493	.000[c]
Interval						
Ordinal	by	Spearman Correlation	.096	.005	21.493	.000[c]
Ordinal						
N of Valid Cases			49858			

threshold_Yes_No * ema3_threshold_YesNo Crosstabulation

		ema3_threshold_YesNo		
		0	1	Total
threshold_Yes_No	0	35006	3487	38493
	1	10082	1283	11365
Total		45088	4770	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	50.445[a]	1	.000		
Continuity Correction[b]	50.187	1	.000		
Likelihood Ratio	48.715	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	50.444	1	.000		
N of Valid Cases	49858				

Symmetric Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	by	Pearson's R	.032	.005	7.106	.000[c]
Ordinal by Ordinal	by	Spearman Correlation	.032	.005	7.106	.000[c]
N of Valid Cases			49858			

threshold_Yes_No * ema5_threshold_YesNo Crosstabulation

		ema5_threshold_YesNo		
		0	1	Total
threshold_Yes_No	0	34578	3915	38493
	1	10175	1190	11365
Total		44753	5105	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.860[a]	1	.354		
Continuity Correction[b]	.827	1	.363		
Likelihood Ratio	.856	1	.355		
Fisher's Exact Test				.360	.182
Linear-by-Linear Association	.860	1	.354		
N of Valid Cases	49858				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 1163.67.

b. Computed only for a 2x2 table

Symmetric Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	by	Pearson's R	.004	.005	.927	.354[c]
Ordinal by Ordinal	by	Spearman Correlation	.004	.005	.927	.354[c]
N of Valid Cases			49858			

Appendix G

threshold_Recoded * sma3_threshold Crosstabulation

		sma3_threshold				
		<5%	5-10%	10-15%	15%	Total
threshold_Recoded	<5	36807	1039	291	356	38493
	Y	4702	306	62	127	5197
	A	2182	259	49	65	2555
	R	2994	126	85	408	3613
Total		46685	1730	487	956	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2547.308[a]	9	.000
Likelihood Ratio	1562.973	9	.000
Linear-by-Linear Association	1855.383	1	.000
N of Valid Cases	49858		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 24.96.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.193	.007	43.898	.000[c]
Ordinal by Ordinal	Spearman Correlation	.160	.006	36.086	.000[c]
N of Valid Cases		49858			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

threshold_Recoded * sma5_threshold Crosstabulation

		sma5_threshold				
		less than 5%	5-10%	10-15%	15%	Total
threshold_Recoded	<5	36432	1300	366	395	38493
	Y	4677	274	100	146	5197
	A	2214	183	64	94	2555
	R	2932	84	97	500	3613
Total		46255	1841	627	1135	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2780.888[a]	9	.000
Likelihood Ratio	1655.006	9	.000
Linear-by-Linear Association	2038.913	1	.000
N of Valid Cases	49858		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 32.13.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.202	.007	46.106	.000[c]
Ordinal by Ordinal	Spearman Correlation	.146	.006	32.914	.000[c]
N of Valid Cases		49858			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

threshold_Recoded * wma1_threshold Crosstabulation

		wma1_threshold				
		less than 5%	5-10%	10-15%	15%	Total
threshold_Recoded	<5	37770	329	184	210	38493
	Y	4874	174	79	70	5197
	A	2425	64	33	33	2555
	R	3253	157	58	145	3613
Total		48322	724	354	458	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1060.195[a]	9	.000
Likelihood Ratio	779.747	9	.000
Linear-by-Linear Association	783.966	1	.000
N of Valid Cases	49858		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 18.14.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.125	.007	28.222	.000[c]
Ordinal by Ordinal	Spearman Correlation	.132	.006	29.820	.000[c]
N of Valid Cases		49858			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

threshold_Recoded * wma3_threshold Crosstabulation

		wma3_threshold				
		less than 5%	5-10%	10-15%	15%	Total
threshold_Recoded	<5	37332	644	206	311	38493
	Y	4757	258	83	99	5197
	A	2355	110	37	53	2555
	R	3094	126	78	315	3613
Total		47538	1138	404	778	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1856.282[a]	9	.000
Likelihood Ratio	1187.552	9	.000
Linear-by-Linear Association	1459.724	1	.000
N of Valid Cases	49858		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 20.70.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.171	.007	38.778	.000[c]
Ordinal by Ordinal	Spearman Correlation	.150	.006	33.926	.000[c]
N of Valid Cases		49858			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

threshold_Recoded * wma5_threshold Crosstabulation

		wma5_threshold				
		less than 5%	5-10%	10-15%	15%	Total
threshold_Recoded	<5	36941	963	253	336	38493
	Y	4728	288	66	115	5197
	A	2256	191	40	68	2555
	R	3015	107	99	392	3613
Total		46940	1549	458	911	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2367.292[a]	9	.000
Likelihood Ratio	1443.309	9	.000
Linear-by-Linear Association	1811.735	1	.000
N of Valid Cases	49858		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 23.47.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.191	.007	43.359	.000[c]
Ordinal by Ordinal	Spearman Correlation	.153	.006	34.506	.000[c]
N of Valid Cases		49858			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

threshold_Recoded * ema1_threshold Crosstabulation

		ema1_threshold				
		less than 5%	5-10%	10-15%	15%	Total
threshold_Recoded	<5	36925	904	236	428	38493
	Y	4804	178	42	173	5197
	A	2325	85	36	109	2555
	R	3195	127	64	227	3613
Total		47249	1294	378	937	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	770.471[a]	9	.000
Likelihood Ratio	598.815	9	.000
Linear-by-Linear Association	740.092	1	.000
N of Valid Cases	49858		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 19.37.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.122	.006	27.409	.000[c]
Ordinal by Ordinal	Spearman Correlation	.102	.005	22.824	.000[c]
N of Valid Cases		49858			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

threshold_Recoded * ema3_threshold Crosstabulation

		ema3_threshold				
		less than 5%	5-10%	10-15%	15%	Total
threshold_Recoded	<5	35006	2641	269	577	38493
	Y	4623	302	70	202	5197
	A	2295	88	32	140	2555
	R	3164	82	69	298	3613
Total		45088	3113	440	1217	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1009.042[a]	9	.000
Likelihood Ratio	829.621	9	.000
Linear-by-Linear Association	458.685	1	.000
N of Valid Cases	49858		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 22.55.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.096	.006	21.516	.000[c]
Ordinal by Ordinal	Spearman Correlation	.039	.005	8.691	.000[c]
N of Valid Cases		49858			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

threshold_Recoded * ema5_threshold Crosstabulation

		ema5_threshold				
		less than 5%	5-10%	10-15%	15%	Total
threshold_Recoded	<5	34578	2765	467	683	38493
	Y	4734	150	125	188	5197
	A	2323	40	26	166	2555
	R	3118	117	55	323	3613
Total		44753	3072	673	1360	49858

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1138.311[a]	9	.000
Likelihood Ratio	984.008	9	.000
Linear-by-Linear Association	349.913	1	.000
N of Valid Cases	49858		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 34.49.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.084	.006	18.772	.000[c]
Ordinal by Ordinal	Spearman Correlation	.016	.005	3.517	.000[c]
N of Valid Cases		49858			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.