Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2010 Proceedings

Americas Conference on Information Systems (AMCIS)

8-2010

Using Text Mining to Analyze Quality Aspects of Unstructured Data: A Case Study for "stock-touting" Spam Emails

Mohamed Zaki
University of Manchester, Mohamed.Zaki@postgrad.mbs.ac.uk

David Diaz
University of Manchester, David.DiazSolis@postgrad.mbs.ac.uk

Babis Theodoulidis
University of Manchester, B.Theodoulidis@mbs.ac.uk

Using Text Mining to Analyze Quality Aspects of Unstructured Data: A Case Study for "stock-touting" Spam Emails

Mohamed Zaki
Manchester Business School
University of Manchester
Manchester M13 9SS
United Kingdom
Mohamed.Zaki@postgrad.mbs.ac.uk

David Diaz
Manchester Business School
University of Manchester
Manchester M13 9SS
United Kingdom
David.DiazSolis@postgrad.mbs.ac.uk

Babis Theodoulidis
Manchester Business School
University of Manchester
Manchester M13 9SS
United Kingdom
B.Theodoulidis@mbs.ac.uk

ABSTRACT

The growth in the utilization of text mining tools and techniques in the last decade has been primarily driven by the increase in the sheer volume of unstructured texts and the need to extract useful and more importantly, quality information from them. The impetus to analyse unstructured data efficiently and effectively as part of the decision making processes within an organization has further motivated the need to better understand how to use text mining tools and techniques.

This paper describes a case study of a stock spam e-mail architecture that demonstrates the process of refining linguistic resources to extract relevant, high quality information including stock profile, financial key words, stock and company news (positive/negative), and compound phrases from stock spam e-mails. The context of such a study is to identify high quality information patterns that can be used to support relevant authorities in detecting and analyzing fraudulent activities.

Keywords

Text Mining, Unstructured Data, Data and Information Quality, Spam Emails.

INTRODUCTION

The rapid development and penetration of technology and communications in our society has strongly affected the flow of information, often false and low quality, on the stock markets. In fact, spam e-mails have been used extensively and creatively to target specific stocks in order for fraudsters to gain illegal benefits. Most commonly, spammers state that they have ascertained private information about stocks. The e-mails contain fine print messages claiming valuable information such as investment advice, stimulation about a specific investment decision disclosed with financial terms and recent price quotes. Thus, stock spammers speculate on positive price models of the traded stocks and sending thousands of e-mails to possible investors to drive the price of the touted stock upwards or downwards.

Unsolicited stock spam emails have been used extensively and creatively. The internet provides an excellent basis for spammers to implement their ideas quickly and cheaply due to the presence of a large number of traders who respond to news

and act on recommendations (Hanke and Hauser, 2008). Thus, financial markets are vulnerable to false information leading to "pump and dump manipulation" (Leinweber and Madhavan, 2001).

Earlier work mentions that over 80% of all e-mail traffic is classified as spam e-mails, with 15% of these messages related to stock touts (Frieder and Zittrain, 2007). Given the fact that the volume of stock spam e-mails is massive, they can be considered one of the main unstructured information quality problems. Previous work indicated that manual classification and filtration financial analysis techniques based on regression models, time-series, and pooled OLS techniques have been successfully used and they concluded that touted spam works in the stock market and there are recipients who act upon it (Boehme and Holz, 2006). However, these techniques are very time and human resource intensive and error-prone and they are very ineffective in dealing with the massive volume of spam e-mails. Furthermore, the most common type of analysis looks at a limited set of email features with implications on the quality of the analysis results. For example, attributes like time stamp, stock profile, symbol tickers, projection, positive and negative news have not considered.

In this context, this paper proposes the use of text mining techniques to improve stock spam e-mail classification and categorization. More specifically, the paper introduces a text mining approach for the analysis of "stock-touting" spam e-mails and discusses the development of a stock spam email architecture considering the various attributes and the quality characteristics of the information contained in spam e-mails.

Previous work

Many previous research papers have identified spam e-mails as an important problem in stock exchanges that need further investigation. For example, Boehme and Holz (2006) found that there is evidence of the harmful effect of the spam messages on the financial markets. The paper datasets were downloaded from Richardson's Stock Spam Effectiveness Monitor (SSEM) archive (Richardson, 2005). The research analyzed 21,935 stock spam messages between November 2004 and February 2006. On average, 3% of all incoming messages were classified as stock spam (Richardson, 2005).The spam messages advertised 391(68% of stock listed in Pink Sheets, 5% listed in OTC market) unique stocks. The study employed a multiplicative multivariate regression model and a classical event study methodology, focusing on the effects of spam e-mails on the return and volume of the target stocks. The research found indications of an increase in trading activity of the advertised stock. Moreover, there was evidence of abnormal returns occurred after the messages have been disseminated. Stock prices found to respond positively to spam e-mails, with a positive relationship between the amount of spam mails per day and the size of returns (Hanke and Hauser, 2008).

Frieder and Zittrain (2007) evaluated and analyzed the impact of touted spam on the trading activity of specific stocks using a big sample from the Pink Sheets market. The researcher used a specific touted stock which is listed on the Pink Sheets quotation system and a sample of spam e-mails. The initial dataset consisted of a database of 1,802,016 unsorted spam messages, most of which were downloaded from the internet usenet newsgroup "news.asmin.net-abuse.sightings (NANAS)". The authors manually extracted the stock tout messages by selecting the ones that contain the word 'stock' and a symbol ticker. This extraction process filtered 75,415 messages with 28,803 different stock symbols between 22/08/00 and 02/08/05. In addition, it filtered down 3,669 symbol date groups with 500 distinct stocks that were touted. The researcher compared the touted stock with another control sample (not touted) during the same period (Hanke and Hauser, 2008). The paper matched the volume and price data from Pink Sheets market with the ticker symbols that has been cited in e-mails to compare the presence and volume of spam that touts particular stocks with the price and volume behaviour of those stocks before, during, and after the touts. Evidence was found of a significant positive return on days that have a heavy spam touting. Furthermore, the volume of trading was corresponding positively to the heavy spam touting. For example, the stock that was touted through spam, it was probably the most active stock in trading operations. It jumped from 4% to 70% in terms of daily return on the day with more touting activity. Moreover, the returns of the following days were significantly negative, confirming the assumption that spammers "buy low and spam high".

Hanke and Hauser (2008) started the analysis by describing the common characteristics of the advertised stock such as price level and average turnover. In spite of investigating the effect of touting spam on returns and volume only, the researcher measured the effect of spam email on other variables like excess returns, turnover and intra-day volatility. The data sample of this study has been collected from the "Crummy database". Partially, it overlaps with the sample used by Frieder and Zittrain (2007) with regard to both ticker symbols and time frame. The study employed panel regression allowing for fixed cross-section effects focused on a sample of spammed stocks. The argument of the paper is that there is a significant positive impact on the securities prices when the touting occurs. Furthermore, the liquidity is the one of the foremost factor in the spamming campaign success. Lastly, the repetitive spam on consecutive days continued to rise up the demand on the targeted stock which strength the spammer position and enlarge their time window for liquidation.

In terms of spam e-mail literature, there are extensive research utilizing various text mining algorithms to filter spam e-mails like (Sahami et al., 1998; Drucker et al., 1999; Hidalgo et al., 2000; Secker et al., 2003; Oda and White, 2005; Ahmed and Mithun, 2004; Kong et al., 2006; Taylor, 2006; Li and Hsieh, 2006; Goodman and Yih, 2006; Surendran at al., 2005; Fumera et al., 2005; Richard et al., 1999). Mainly these studies discussed the use of two filtering methods; the first method is blocking the sources of spam based on various features like e-mail address and e-mail servers. The second method is analyzing the content of spam e-mails by employing the rule-based method. Content-based filtering is utilizing text mining techniques and machine learning approaches like Naïve Bayes and Support Vector Machine (SVM) to classify models and generate automatic content filtering rules. Basically, these rules are analyzing words and phrases considering its appearance and distribution in the e-mails context (Yeh et al., 2007).

In related applications of text mining techniques, some studies employed text mining techniques to analyze the news contents and measure its effect on the financial market. For example, Fung et al. (2002) and (2003) propose a methodology to evaluate the immediate impact of news article on the stock trend prediction. Trend labelling clusters were used to segment trends into two categories up or down. Mittermayer (2004) developed News categorization and Trading System (NewsCATs) which could predict stock prices after news release. Support Vector Machine (SVM) text classifier has been used for the categorization process which basically flag good or bad news. Other studies like (Schumaker and Chen, 2006; Khare et al., 2004; Failinouss (2007) proposed an automatic text extraction process by utilizing different classifiers like Naïve classifier or SVM to predict stock price changes based on analyzing the effect of news article. The main methodology for these studies is extracting key terms, analysing headline and paragraph, sentiment classification, and finally stock market prediction.

The aforementioned studies showed that touted spam is working within the stock market and there are recipients that read and also act upon it. However, the analysis and evaluation of the previous studies are based on manual collection and analysis of e-mails. Furthermore, most of the studies are focusing on one or two features of the e-mails like symbol ticker, and tout volume neglecting the possibility of more complete analysis, such as a content based analysis or sentiment analysis. In this sense, this paper contributes to previous research by providing an alternative portfolio of methods and techniques based on text mining that could be leveraged in the process of detecting information-based manipulations in the financial spam e-mail context.

Moreover, the research also focuses on building a conceptual architecture for the financial spam e-mail study, extending the analysis to develop patterns that can refine linguistic resources to extract financial key news, company news, concept, terms and compound phrases. In addition, it helps in identifying the efficient patterns to undertake the information-based manipulation problem with a unified body of knowledge that could support relevant authorities in detecting fraudulent activities.

Expected outcomes of the research are a reliable stock spam e-mail taxonomy considering the content analysis of these e-mails, for instance, extracting common spammer words, phrases, and sentiments enclosed with e-mails speculating positive or negative news. Also, this work will contribute to reveal possible classifications, clusters that could contribute in showing the influence of specific message draft characteristics on its market impact. In addition, it could yield interesting findings which persuasive features enclosed in these messages that heavily increase the response rates of spammer's campaigns.

Stock Spam E-mail Architecture

Spam emails are collected on various locations (blogs, newsgroups, websites) using trap email accounts. There are also a number of websites that advice and alert users and also, archive and classify spam messages according to their date and time (Hanke and Hauser, 2008). In this section, the proposed stock spam e-email architecture is discussed (see figure 1).

The proposed architecture assumes that the financial data are from the Pink Sheets stock market and it also assumes that data are analysed after the actual transactions are finalised (no real-time features, retroactive analysis, limited input sources) (Diaz et al., 2010b). Before any analysis take place, the data need to be pre-processed (Data Preparation) in order to combine and match different sources and types of data. Spam e-mails are a very good example of very dirty unstructured data and during this process a lot of work are dedicated to improving the quality and addressing any inconsistencies, inaccuracies and omissions.

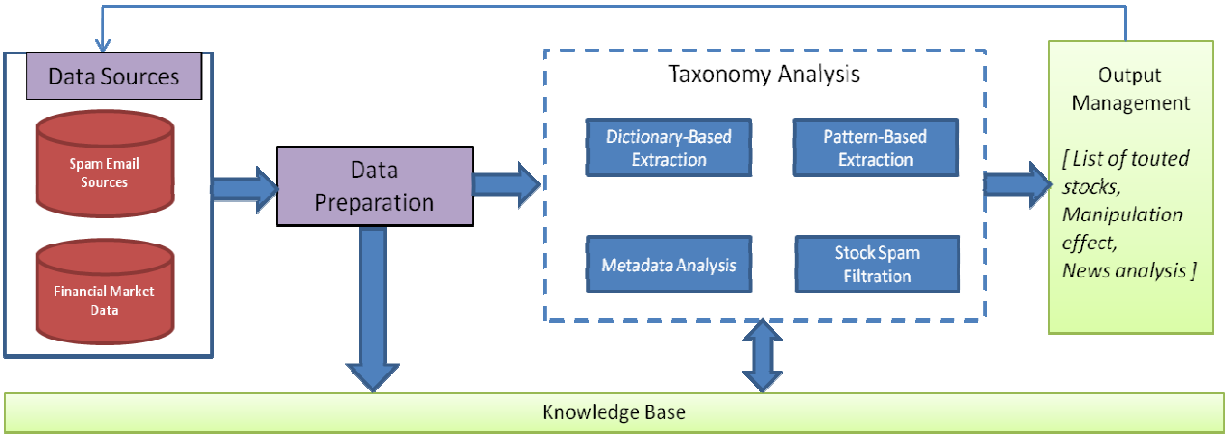


Figure 1. Stock Spam E-mail Architecture

Two methods for the extraction process are proposed (see Figure 1), namely, dictionary-based and pattern-based. The dictionary-based extraction method requires configuring and training systems with linguistic resources like thesaurus, list of terms, concepts, synonyms, types (semantic groupings of concepts). For example, the dictionary could have a list of common financial terms that are heavily cited in the stock e-mails and a list of traded stock symbol tickers collected from the market. In addition, the dictionary could have an exclude list of terms that are not useful for the analysis. This method could use tokenization and morphological analysis to extract terms from e-mails and match them with the appropriate entities. Tokenization analysis parses documents into characters and words which are called tokens. Text mining systems are using advanced algorithms that deal with tokenization challenges like punctuation, dash symbols, apostrophe character and others. For the morphological analysis, stemming algorithms are employed to identify the root of terms listed in the dictionary. Text mining applications integrate stemming algorithms on the tokenization output to conflate the tokens into an orthogonal set of distinct morphological groups which are used to train the extraction engine to group similar forms (singular or plural) of terms and add them to the dictionary. However, this method is time consuming and can produce inaccuracies if the terms are not appropriate for the financial analysis domain (Froelich and Ananyan, 2008).

The pattern-based extraction method complements the dictionary-based method by helping to define relationships between extracted concepts in the e-mail messages based on natural language pattern rules. This method performs a deeper analysis on the words, phrases and syntax inside the e-mails and thus, it helps to uncover knowledge of the underlying language and these e-mails messages. The method offers higher degree of accuracy and it could classify whether the news cited in the e-mail are positive or negative. It incorporates a named entity recognition analysis which identifies and extracts certain non-linguistic terms and maps them to real-world entities or events such as person, e-mail address, postal address, etc. Moreover, the pattern-based method utilises word association analysis to identify the relationship between two or more entities which could lead to a better understanding of the way that these relationships could affect stock prices or volume movement (upward or downward) in the market.

The proposed architecture identifies the importance of analyzing the metadata associated with any e-mail. These include information about the date and time the e-mails were sent, elements like "from" or "to" that stores information about name or the e-mail address of the sender or the receiver, etc. The subject is another interesting element to be analyzed as it could be one of the attraction features to tempt investors to open and read the e-mail. Overall, the metadata of spam e-mails are useful to analyse but they could contain misleading and incorrect information. The time stamp analysis is a very useful way to analyse for example, when spammers start their campaign for a specific stock, and how long spammers continue their campaigns for specific stocks. Thus, each metadata element could generate interesting findings to help the relevant authorities to build a better analysis on how stock spam emails work in the financial market.

Furthermore, the architecture includes a filtration classifier (stock spam filtration) to assign documents to known categories like classifying stock spam documents from ordinary spam e-mails. For example, the system could be uploaded with a list of stock symbols (Dictionary) for specific market like pink sheets and has rules which find keywords like symbol, stock, and others that could distinguish stock spam e-mails from others. The result of this process exports the identified stock spam e-mails in the knowledge base.

Finally, the architecture generates the required taxonomy for stock spam e-mails, annotates interesting concepts and explains the relations between entities. The taxonomy is based on two types of analysis as shown in table 1. The metadata analysis deals with key header information like date, time, sender, receiver, and subject. The body analysis deals with extracting named entities of stock spam e-mails for further analysis. For example, it extracts information on stock profiles such as the symbol ticker, the company holding the stock, and the sector that the company is related, and also, stock speculations that could tempt investors such as price or volume projections, trading date expectations, recommendations, and financial investment indicators either long or short term, also any financial signals like buying signal. Furthermore, the architecture aims to extract the concepts and phrases that indicate whether the cited news is related to the stock (e.g. trading description) or the company (e.g. agreements) itself. This feature could give some positive or negative feelings to investors and encourage them to respond and trade following the recommendations that have been given by spammers.

Metadata Analysis	**E-mail Features**
Header Analysis	• Date and Time "Date" • Sender "From" • Receiver "To" • E-mail Subject "Subject"
Body Analysis	**E-mail Features**
Stock Profile	• Stock Symbol Ticker • Company name (stock holder) • Sector
Stock Speculations	• Stock price speculation • Volume Speculation • Financial investment indicators (long/short terms) • Buy signal indicators • Trade date expectation
News	• Stock News (positive/negative) • Company news (positive/negative)
Recommendations	• Positive • Negative

Table 1. Stock spam email taxonomy output

Stock Symbol Ticker Extraction Case

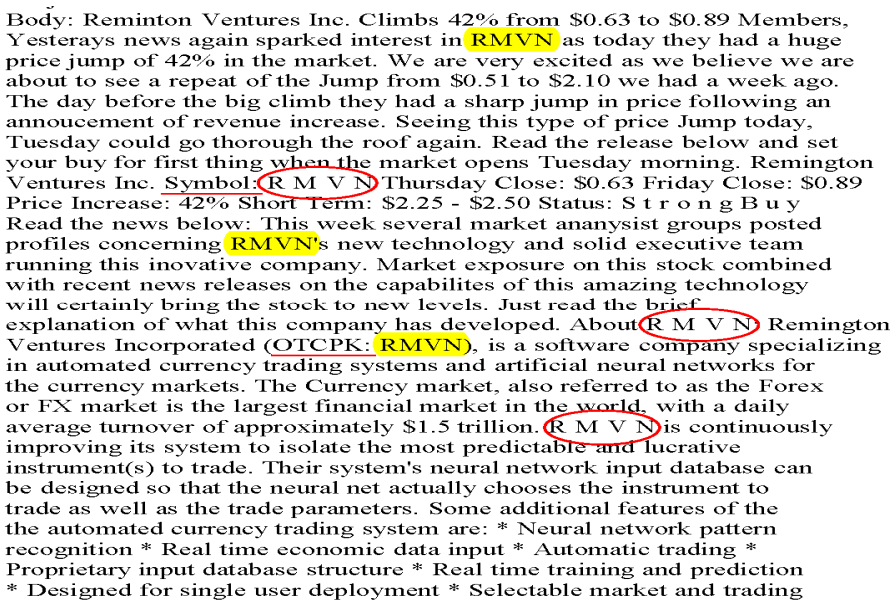
This section demonstrates an example of the symbol ticker extraction process by adapting the stock spam e-mail architecture. This work will follow the architecture recommendation by balancing two extraction methods (Dictionary and Pattern based analysis) to generate better extraction results for the symbol tickers. This study has performed and ran using Text mining analysis tools from SPSS Clementine software.

Specifically, the spam e-mails dataset used in this study was collected from the Richardson's Stock Spam Effectiveness Monitor (SSEM) archive (Richardson, 2007). The data source contains messages from January 2006 to February 2008. Initially, the spam e-mail messages were archived using a UNIX/LINUX format (MBOX format), but later on messages were converted to HTML standard text format. Figure 2 show an example of one stock spam e-mail from this dataset.

In addition, the study used market data collected from the Pink Sheets market. The data comprises an updated list (October 2009) of 23,028 over the counter (OTC) traded stock (Symbol ticker, company name). Moreover, Pink Sheets market provided the researcher with a list of issued symbol ticker and company name from 2006 to October 2009. Regarding the market data, the data preparation phase covered appending and sorting activities to construct the final dataset, i.e., data that was fed into the analysis from the initial raw data. The OTC and Pink Sheets data sources were appended to create one list of issued stock symbol ticker and company list to be used

in the analysis phase. Restructured market data, along with converted e-mails documents are stored in a knowledge base to feed the taxonomy analysis with a clean dataset.

In terms of analysis, initially restructured market data was used to build a dictionary for stock symbol ticker. The dictionary has 25,295 symbol tickers. The symbol ticker contains from 4-5 characters. During the extraction process, the e-mails were scanned and analyzed in order to identify symbol tickers words within the dictionary and map the symbols to an event called ***normal cases***. In Figure 2 it is possible to appreciate how one of these stock symbols is mapped by highlighting the acronym (RMVN). Given the fact that these symbol tickers are issued without consideration of the lexical meaning, some of these symbols create conflicts with English words. For instance, words like "every", "this", "auto", are listed as stock symbols. This could lead to conflict in the extraction process as many of these words will be mapped to ***normalcases*** event. In order to overcome this conflict, an alternative dictionary set was built mapping these cases into another event called ***specialcases***.



Body: Reminton Ventures Inc. Climbs 42% from $0.63 to $0.89 Members, Yesterays news again sparked interest in RMVN as today they had a huge price jump of 42% in the market. We are very excited as we believe we are about to see a repeat of the Jump from $0.51 to $2.10 we had a week ago. The day before the big climb they had a sharp jump in price following an annoucement of revenue increase. Seeing this type of price Jump today, Tuesday could go thorough the roof again. Read the release below and set your buy for first thing when the market opens Tuesday morning. Remington Ventures Inc. Symbol: R M V N Thursday Close: $0.63 Friday Close: $0.89 Price Increase: 42% Short Term: $2.25 - $2.50 Status: S t r o n g B u y Read the news below: This week several market ananysist groups posted profiles concerning RMVN's new technology and solid executive team running this inovative company. Market exposure on this stock combined with recent news releases on the capabilites of this amazing technology will certainly bring the stock to new levels. Just read the brief explanation of what this company has developed. About R M V N Remington Ventures Incorporated (OTCPK: RMVN), is a software company specializing in automated currency trading systems and artificial neural networks for the currency markets. The Currency market, also referred to as the Forex or FX market is the largest financial market in the world, with a daily average turnover of approximately $1.5 trillion. R M V N is continuously improving its system to isolate the most predictable and lucrative instrument(s) to trade. Their system's neural network input database can be designed so that the neural net actually chooses the instrument to trade as well as the trade parameters. Some additional features of the the automated currency trading system are: * Neural network pattern recognition * Real time economic data input * Automatic trading * Proprietary input database structure * Real time training and prediction * Designed for single user deployment * Selectable market and trading

Figure 2. Example for the stock spam e-mail

However, there is evidence that spammers are smart and know that some of this dictionary based filtration systems are already in place. Consequently, they use a wide range of tactics to deceive this type of automatic filtration systems. For example, the symbol ticker RMVN could be cited in different ways like R M V N, R-M-V-N, RMVN.PK, or other similar tactics which do not work under the symbol dictionary method. Therefore, the dictionary-based method is not effective enough for the task.

Pattern-based analysis could help addressing these tactics by considering a different approach. In this study, three patterns are defined. Particularly, each pattern can be understood as a set of rules that describe structures in words, combinations of them or combinations of rules. For example, if the task is to find or recognize symbol tickers been mentioned in the e-mail body or text, it is possible to define a rule in which all collection of 4-5 capital characters are mapped to an entity called ***specialpatterns***. An advantage of this approach is that is possible to extract not only stocks tickers that were originally listed in the dictionary, but also to extract new issued symbol tickers that matched the predefined rules.

In the same line, a second pattern called ***symbolkeyword*** was defined in order to extract the word "symbol". This is because it is been noticed that commonly the word symbol is preceded by a symbol ticker, for example, "Symbol: RMVN". Additionally, different forms of the word "symbol" has been considered, such as, "stock symbol", "symb", "sym", "etssymbol", "o.t.c. sym bol", "o.t.c symb0l", and others.

As shown in figure 2, there is another possibility that the stock exchange name or market venue may precede the symbol ticker, such as, "OTCPK: RMVN". Again, another pattern called ***MKTSymbol*** was developed to resolve

this issue, also considering different combination of the market names like "otcpk", "otc","otc.bb", "otcbb", "other otc", and others.

By balancing the two extraction methods, the output of this analysis is a list of touted stock symbols from stock spam e-email messages from the original dataset. Table 2 shows some examples of cited symbol ticker and how the two methods were employed to map the symbol tickers to its entities.

CONCLUSION

Spam emails have been used as one of the main information-based manipulation tools utilising various techniques to disseminate false, incomplete or inaccurate information to potential traders. The stock spam e-mail architecture introduced in this paper utilises a text mining approach for the analysis of "stock-touting" spam e-mails to improve the analysis of the disseminated information.

The proposed framework could be used as a methodology to correlate the common characteristics of spam e-mails and improve the quality of the extracted information. The results of the analysis could inform potential investors from possible manipulation schemes and also, support analysts and relevant authorities to identify possible touting cases.

Further research will address the extension of the architecture by utilizing data mining techniques to correlate the content and the features of spam e-mail messages with price movement and trading behaviour (Zaki et al., 2010). The author will evaluate the architecture by applying various cases and it will be quantitatively assessed. This work could also incorporate previous work by the authors in applying data mining techniques to detect stock price manipulations through the analysis of intraday trade prices and closing prices for the investigation of trade-based manipulations (Diaz et al., 2010a).Furthermore, financial indicators will be developed to support the data mining techniques to detect any suspicious trading.

Stock Symbol Ticker example	Dictionary-Based Analysis	Pattern-Based Analysis
RMVN	Normal cases	
R M V N\ R.M.V.N\ R-M-V-N.pk and others		<SpecialPatterns>
OTCPK.RMVN	Normal cases	<MKTSymbol>
Symbol: RMVN	Normal Cases	<SymbolKeword>
Symbol: R_M_V_N		<SymbolKeword > < SpecialPatterns>
(New issued symbol) Symbol: ESTE		<SymbolKeword > < SpecialPatterns>
Symbol: THIS	Special Cases	<SymbolKeword>

Table 2. Stock symbol ticker extraction analysis output

REFERENCES

1. Ahmed, S., & Mithun, F. (July 2004). Word Stemming to Enhance Spam Filtering. *In Proceedings of the First Conference on Email and Anti-Spam.*
2. Böhme, R., & Holz, T. (2006). The Effect of Stock Spam on Financial Markets, Working Paper. *Available at SSRN: http://ssrn.com/abstract=897431.*
3. Diaz, David, Theodoulidis, Babis and Sampaio, Pedro, Analysis of Stock Market Manipulations Using Knowledge Discovery Techniques Applied to Intraday Trade Prices (March 1, 2010). Available at SSRN: http://ssrn.com/abstract=1561882
4. Diaz, David, Theodoulidis, Babis and Sampaio, Pedro, A Market Manipulation Monitoring System Framework (March 1, 2010). Available at SSRN: http://ssrn.com/abstract=1561663

5. DomainKeys.(n.d.).*Proving and Protecting Email Sender Identity.* Nuskaityta is: http://antispam.yahoo.com/domainkeys
6. Drucker, H. D., Wu, D., & Vapnik., V. (1999). Support Vector Machines for spam categorization., . *IEEE Trans. on Neural Networks, 10(5):1048-1054* .
7. F. li, & Hsieh., M. H. (July 2006). An Empirical Study of Clustering Behavior of Spammers and Group-based Anti-Spam Strategies. *In Proceedings of the Third Conference on Email and Anti-spam.*
8. Falinouss, P. (2007). Stock Trend predction using news articles. Lulea university of technology. Lulea University of Technology.
9. Frieder, L. L., & Zittrain, J. (2007). Spam Works: Evidence from Stock Touts and Corresponding Market Activity. *Available at SSRN: http://ssrn.com/abstract=920553*.
10. Froelich, J., & Ananyan, S. (2008). *Decision Support via Text Mining.* Springer Berlin Heidelberg.
11. Fumera, G., Pillai, I., & Roli., F. (December 2006). Spam Filtering Based On The Analysis Of Text Information Embedded Into Images. *Journal of Machine Learning Research, 7:2699-2720* .
12. Fung, G. h., Yu, J. X., & Lam, W. (2002). News Senestive stock trend Prediction . *Pacafic Assia Conference (PAKDD) on Advanced in Knowledge Discovery and data Mining* , (psl. 481-493). Tiapei.
13. Fung, G. h., Yu, J. X., & Lam, W. (2003). Stock Prediction: Integrating Text Mining Approach using Real-Time News. *IEEE International Conference on Computational intelligence for financial Engineering (CIFER)* (psl. 395-402). Hong Kong (China): IEEE Press .
14. Fung, G. P., Yuy, J. X., & Luz, H. (2005). The Predicting Power of Textual Information on Financial Markets. *IEEE Intelligent Informatics Bulletin Vol.5 No.1* .
15. Goodman, J., & Yih, W. T. (July 2006). Online Discriminative Spam Filter Training. *In Proceedings of the Third Conference on Email and Anti-spam.*
16. Hanke, M., & Hauser, F. (2008). On the effects of Stock Spam Emails. *Journal of Financial markets 11* , 57-83.
17. Hidalgo, J. G., & Lopez., M. M. (2000). Combining text and heuristics for cost-sensitive spam filtering. *Computational Natural Language Learning Workshop* , 99-102.
18. Kong, J. S., Rezaei, B. A., Sarshar, N., Roychowdhury, V. P., & Boykin, P. O. (August 2006). Collaborative Spam Filtering Using Email Networks. *Computer, 39(8)* , 67-73.
19. Leinweber, D. J., & Madhavan, A. N. (2001). Three Hundred Years of Stock Market Manipulations. JOURNAL OF INVESTING .
20. Oda, T., & White, T. (August 2005). Immunity from Spam: An Analysis of an Artificial Immune System for Junk Email Detection. *The 4th International Conference on Artificial Immune Systems*, (psl. 276-289).
21. R. Khare, N. P. (2004). Stock Broker P – sentiment extraction for the stock market. Data Mining V, proceedings of the Fifth International Conference on Data Mining, Text Mining and Their Business Applications (psl. 15-17). Malaga, Spain: WIT Press.
22. Richard, B. S., & Jeffrey, O. K. (1999). MailCat: an intelligent assistant for organizing e-mail. *The third annual conference on Autonomous Agents*, (psl. 276 - 282).
23. Richardson, L. (2007 m. August 17 d.). *Stock spam effectiveness monitor*. Paimta 2008 m. August 9 d. iš http://www.crummy.com/features/StockSpam/
24. Sahami, M., Dumais, S., Heckerman, D., & Horvitz, a. E. (55-62). ABayesian approach to filtering junk e-mail. . *AAAI Workshop on Learning for Text Categorization* , 1998.
25. Schumaker, R. P., & Chen, H. (2006). Textual analysis of stock market prediction using financial news. *Americas Conference on Information Systems.* Mexico: Acapulco.
26. Secker, A., Freitas, A. A., & Timmis, J. (2003). AISEC: An Artificial Immune System for Email Classification. *The IEEE Congress on Evolutionary Computation Proceedings, 1* , 131-138.
27. Surendran, A. C., Platt, J. C., & Renshaw., E. Automatic Discovery of Personal Topics to Organize Email. *In Proceedings of the Second Conference on Email and Anti-Spam.* July 2005.
28. Taylor, B. (July 2006). Sender Reputation in a Large Webmail Service. *In Proceedings of the Third Conference on Email and Antispam.*

29. Wong, M., & Schlitt., W. (n.d.). *Sender Policy Framework (SPF) for Authorizing Use of Domains in E-mail.* Nuskaityta iš http://www.openspf.org/Project_Overview

30. Yeh, C.-F., Mao, C.-H., Lee, H.-M., & Chen, T. (2007). Adaptive E-mail Intention Finding Mechanism based on E-mail Words Social Networks. *Applications, Technologies, Architectures, and Protocols for Computer Communication archive* (psl. 113-120). New York, NY, USA: ACM .

31. Zaki, Mohamed, Theodoulidis, Babis and Diaz, David, A Data Mining Approach for the Analysis of "Stock-Touting" Spam Emails (February 15, 2010). Available at SSRN: http://ssrn.com/abstract=1561906