A DATA MINING APPROACH FOR THE ANALYSIS OF
“STOCK-TOUTING” SPAM EMAILS

Mohamed Zaki, Babis Theodoulidis, David Diaz Solis
University of Manchester, Manchester Business School, Manchester M13 9SS, United
Kingdom, Mohamed.Zaki@postgrad.mbs.ac.uk, B.Theodoulidis@mbs.ac.uk,
David.DiazSolis@postgrad.mbs.ac.uk

Abstract. Although the financial markets are regulated by robust systems and rules that control their efficiency and
try to protect investors from various manipulation schemes, markets still suffer from many attempts to mislead or
misinform investors in order to gain illegal profits. The impetus to effectively and systematically address such
schemes is presenting many challenges to academia, industry and relevant authorities. This paper proposes the use
of data mining techniques to detect abuses in the stock market and in particular, Information-based manipulations.
These are manipulations that allow rogue traders to gain illegal profits from disseminating false or vague
information to investors through spam emails. The paper proposes a spam fraud analysis and detection framework
using data mining techniques that helps analysts to identify possible touting cases based on spam emails. The
framework employs different techniques such as classification, neural networks and linear regressions. The
application of the framework is demonstrated using data from the Pink Sheets market and the results strongly
suggest that data mining techniques can be used to facilitate fraud investigations originating from spam emails.
The proposed framework and findings of the paper could be used in a retroactive mode to help the relevant authorities
and organisations to identify abnormal behaviors in the stock market. It could also be used in a proactive mode to
warn analysts and stockbrokers of possible cases of market abuse.

Keywords: Business Intelligence, Data Mining, Fraud Detection, Spam Emails.

1 Introduction

Financial markets continue to suffer from attempts to mislead or misinform investors to impact the
decision making process to gain illegal profits despite the increased regulations and control systems. The
impetus to effectively and systematically address such schemes presents a very important challenge to academia,
industry and relevant authorities.

The rapid development and penetration of technology and communications in our society has strongly
affected the flow of information within the stock market. 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 specific investment decision
disclosed with financial terms and recent price quotes. Thus, stock spammers speculated on positive price
models of the traded stocks and sent thousands of e-mails to possible investors to tempt investors, driving the
price up/down upon the touted spam.

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 [3]. Various studies have manually classified and categorized spam e-mails
to evaluate the effect of spam volume tout on the market efficiency. They concluded that touted spam
functionally works in the stock market and there are recipients who act upon it [1][3].

Fraud detection systems paly an important role in investigating, and detecting fraudulent behaviours and
ascertain possible evidence of various financial frauds within different financial markets [7]. Furthermore, these
systems could support financial organizations to proactively detect market abuse (MA) transactions, thus
minimizing the circulation of rumors or illegal information [2].

This paper introduces a data mining approach for the analysis of “stock-touting” spam emails by
proposing a number of financial indicators to use for modelling the problem and applying a data mining
approach to analyse the problem. Different algorithms are examined and their performance is compared in terms
of their classification accuracy.

2 Previous work

Many research areas have delivered the spam e-mails as one of the main problems that need further
investigations. Such as, Boehme and Holz [1] found that there is evidence of harmful effect of the spam
messages on the financial market. They employed a multiplicative multivariate regression model and classical
event study methodology, focusing on effects of spam e-mails on the return and volume of the target stocks.
They found some indications of an increase in trading activity of the advertised stock. Moreover, there were an
evidence of abnormal returns occurred after the messages have been disseminated. Stock prices respond
positively to spam e-mails, with a positive relationship between the amount of spam mails per day and the size of returns [5]. Lastly, they linked this kind of manipulation in prices and volumes to three groups of individuals (spammers, Naïve recipient, Smart recipient).

Frieder and Zittrain [3] evaluated and analyzed the impact of touted spam on the trading activity of specific stocks using a big sample from the Pink Sheets market. They compared the touted stock with another control sample (not touted) during the same period [5]. Evidence was found of a significant positive return on days that have a heavy spam touting. Furthermore, the volume of trading was responding 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 was jumping from 4% to 70% daily return on a day that has touting activity. However, the returns of the following days are negative; which is firmly confirming the assumption that spammers “buy low and spam high”.

Hanke and Hauser [5] started their analysis by describing the common characteristics of advertised stock such as price level and average turnover. In spite of investigating the effect of touting spam on returns and volume only, they measured the effect of spam email on other variables like excess returns, turnover and intra-day volatility. They proved that there is a significant (positively) impact on the securities prices when the touting occur. Furthermore, they found that liquidity is the one of the foremost factor in the spamming campaign success. Lastly, the repeated 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. This paper has some limitations such as, the unavailability of intra-day data and bid-ask quotes. These would give the research the ability to evaluate the potential profits from trading strategies.

All the aforementioned studies proofed that touted spam is functionally working within the stock market and there are recipients read it and also act upon it. However, the analysis and evaluation of the previous research only used the normal financial, statistical equations and models to proof their assumptions. At the same time, the amount of data in the stock market has been increased exponentially during the last ten years, and it has reached unprecedented levels of data generation, scarcely compared with any other fields. This paper will continue the effort that have been done by the previous research by finding the appropriate data mining techniques to analyze and detect fraudulent activities of the spam scenario.

Through the previous literature there are different data mining techniques that have been utilized to detect fraudulent activities in different domains such as telecommunications stock exchanges, credit-card and insurance companies, banks and securities firms. Such as, Donoho [2] utilized some data mining techniques to detect the early insider trading manipulation scheme before news breaks within the option market. He studied two cases for two companies released some news which indicated negative trading behaviour and mislead investor decision. As well as his research compared the results of the algorithms that have been used in his experiment like C4.5, decision tree, neural network, K-Mean clustering, and logistic regressions. Furthermore, the research concludes that the early detection of insider trading can avoid the market and investors from potential losses.

Kirkland [6] described how the National Association of Securities Dealers (NASD) Inc. used a fraud detection system called Advanced Detection System (ADS) to monitor trades and quotations in the NASDAQ Stock Market. The main objective of the system as explained was to detect and identify any suspected trading behaviour for further investigation. Furthermore, raising the level of surveillance from issue based to firm based patterns and practices. The system used many artificial intelligence (AI) techniques like visualization, data mining techniques (association rules, decision trees), time sequence pattern, and pattern recognition to cover the market surveillance, trading violation and help the regulatory parties to protect the market from any breaches. Furthermore, the system is keeping up to date with modern patterns and new data mining algorithms to adapt with any changes in the market behaviour or any new trading violation.

Goldberg [4] described the Securities Observation, News, Analysis and Regulation system (SONAR) that developed by NASD. The main objective of the system is monitoring the NASDAQ transactions in the stock market to detect and identify any potential insider trading and any falsification of fraud news. The system uses many techniques like data mining, text mining, statistical regression, fuzzy matching, and rule based inference that are used to mine various news wire stories and U.S. Securities and Exchange Commission (SEC) filings. Furthermore, the system evaluates price and volume models for various securities within the market, and generates flag alerts daily for further investigation that could support the prosecution of those who are involved in the manipulation. [2] Commented that SONAR system is a good approach for the regulatory purposes, a reliable detector of “inside trading” before the news break.

The analysis introduced in this paper is based on specific sample data that has been collected and analyzed by [3]. They used a specific touted stock which is listed on the Pink Sheets quotation system and a sample of spam e-mails. Their initial dataset consisted of a database of 1,802,016 unsorted spam messages, most of which were downloaded from the internet usenet newsgroup “news.admin.net-abuse.sightings (NANAS)”. They 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.

This study utilizes the following web site which contains the raw data compiled by [3] “http://cyber.law.harvard.edu/stockspam”. The paper uses the China World Trade Corporation “CWTD” stock as a case study to detect the highest possibilities of fraudulent activities. All spam messages that have the symbol “CWTD” (the target sample stock in this research) received on 13/02/04 were placed into a cluster; and representative messages were sampled to check if the symbol reflected an actual tout or not.

3 Spam Manipulation Indicators

Before performing the analysis, the main symptoms of spam manipulation in the stock market are analysed. This allows to build a suitable set of financial indicators to detect the fraudulent behaviors.

The primary symptom is a jump in the volume of traded stocks, which normally indicates good news for investors. However, it may make the stock prices fall sharply after artificial stock pumping. Figure 1 shows a visual analysis of the average daily trading volume of “CWTD”. In this representation, normal days present smooth trends with dark blue dots. The dot size is reflected by the number of touting. Small dark blue dots are expected on a normal trading day. In contrast, larger dots represented by different colours indicate a significant increase in volume and touting. This can be associated with high probabilities of trading violations and possible pump and dump manipulations. The most clear suspicious behaviour is seen in the days prior to 03/01/2004 in which the “CWTD” stock was highly touted. This significantly affected the trading volume which reached 1,400,000 transactions (red dots).

![Figure 1. Touted Stock Volume Analysis](image)

Another symptom is the jump in the prices of the touted stocks. The quotation transactions available for this study include the closing prices which represent the last price at the closing time. Unfortunately, this paper does not have access to the intra-day price transactions of “CWTD”. Figure 2 represents the average daily trading prices of “CWTD”. Normal movements are expressed using dark blue dots. Any variation in the size of the dots is linked to the number of touting occurring in that trading day. It is clearly seen that small dark blue dots are indicative of a normal trading day. Larger and lighter coloured dots indicate a significant increase in prices and volume. In the absence of any new information in the market, related to the CWTD company, in the form of financial news, or fillings sent to the authorities, these jumps could be related with high probability to fraudulent activities. For example, there is a clear suspicious behavioural transaction around the 02/01/2004 for the “CWTD”. This date shows the stock was highly touted and therefore significantly affecting the stock price and the trading volume as well.

![Figure 2. Touted Stock Price Analysis](image)
Imbalance between the intraday bid and ask number of offers could be another symptom that could be used as an indicator for detecting fraudulent activities. Stock prices should always be affected by any information swirl in the market which could make stock prices fluctuate. If the amounts of bidding offers are greater than the asking number of offers then this indicates good news. However, more asking offers than bidding offers may signify bad news. If these imbalances are not driven by news, either good or bad, then this could indicate the presence of some form of manipulation. Due to this fact, spammers could try to deceive people by stating that they have some kind of insider information and that they should trade and act upon it. Unfortunately, this study only has access to the last bid and ask price of the day. In order to overcome this limitation we build an indicator based on the average of the bid and ask price. This will be further explained in the following sections.

The mentioned data is publicly available in the authors website, from which the whole stock touts table, was downloaded and then filtered down to match the selected tout stock “CWTD” case. Basically, there are 6 main fields representing the symbol ticker of the touted stock, trading date, bid, ask, price, and volume. This table presents the trading transactions of “CWTD” from 09/03/02 to 09/08/05. A second table was also downloaded to extract the number of tout messages which occurred in this stock. This study begins to prepare the data for the analysis by combining the two tables and adding the field “touted” for the sample period.

There are five main indicators that were used in this paper to analyze the dataset. These indicators are based on the symptoms that are described previously and could be classified into three categories: jump in prices, jumps in volume and bid and ask behaviour.

**Category I Price Indicators**

The following are indicators that build on the jump of prices symptoms.

- **Indicator PInd_1 (Average Price Indicator):** it is the ratio of the change in prices for a day to the moving average (MA) of the change of prices of the last 20 days. The formula is based on the principle that daily changes in prices should lay in the vicinity of the 20 days moving average in order to consider them as “normal behaviour”.

  \[ PInd_{-1t} = \frac{\Delta price_t}{MA(20)_t}, \]  

  where,

  \[ \Delta price_t = \frac{P_t - P_{t-1}}{P_{t-1}} \]  

  \[ MA(20)_t = \frac{\sum_{i=1}^{t-18} \Delta price_i}{20} \]

- **Indicator PInd_2 (Std. Deviation Price Indicator):** it is the ratio of the change in prices for a day to the moving standard deviation changes of price of the last 20 days. The formula is based on the principle that daily changes in prices should lie in the vicinity of the 20 days moving standard deviation in order to measure the relative price. It shows how many standard deviations above average (type of z-score).

  \[ PInd_{-2t} = \frac{\Delta price_t}{\sigma(20)_t}, \]  

  where,

  \[ \sigma(20)_t = \sqrt{\frac{\sum_{i=1}^{t-18} (\Delta price_i - MA(20)_t)^2}{20}} \]
Category II Volume Indicators

The following are indicators that are built on the jump of volume symptoms.

- **Indicator VInd_1 (Average Volume Indicator)**: it is the ratio of the volumes for a day to the moving average (MA) of the volumes for the last 20 days. The formula for calculation is based on the principle that daily changes in volumes should lay in the vicinity of the 20 days moving average in order to consider them as “normal behaviour”.

  \[ VInd_{-1} = \frac{\Delta volume_t}{MA(20)_t} \]

  where,

  \[ \Delta volume_t = \frac{V_t - V_{t-1}}{V_{t-1}} \]

  \[ MA(20)_t = \frac{\sum_{i=1}^{18} \Delta volume_i}{20} \]

  With \( V_t \) the total volume of day \( t \), and with \( t = 1, 2..., n=425 \) in the range 02-01-2004 to 08-09-2005.

- **Indicator VInd_2 (Std. Deviation Volume Indicator)**: it is the ratio of the volumes for a day to the moving standard deviation volumes of the last 20 days. The formula for calculation is based on the principle that daily changes in volumes should lay in the vicinity of the 20 days moving standard deviation in order to measure the relative volumes.

  \[ VInd_{-2} = \frac{\Delta volume_t}{\sigma(20)_t} \]

  \[ \sigma(20)_t = \sqrt{\frac{\sum_{i=1}^{18} (\Delta volume_i - MA(20)_t)^2}{20}} \]

Category III Bids and Asks behaviour Indicators

The following is the indicator that builds on imbalance between bids/asks symptoms.

- **Indicator BAInd_1 (Bids/Asks Indicator)**: it is the average of buying and selling pressures implicit in the bids and asks prices. To build the indicator first, it is necessary to signal a day’s close as a “buy” or “sell” day in terms of how close was the closing price to the bid or the ask closing offers. It is argued that if the closing price of the day is close to the ask price, the buying pressure was predominant that day, because buyers choose to pay this ask price according to their expectations of prices going up in the future. In contrast, if the closing price is close to the bid price, then the selling pressure was predominant that day, because sellers choose to sell at this price according to their expectations of lower prices in the future. On average it could be expected that buying and selling pressures compensates each other, making the price fluctuate around his fundamental value. If the buy or sell pressure is persistent without any fundamental reason or news, this could indicate the presence of a manipulation in the market.

  First to signal a day as a “buy” or “sell” day, a variable is constructed as following:

  \[ BA\_var1 = P_t - A(bids_t, asks_t) \]

  where,

  \[ A(bids_t, asks_t) = \frac{bids_t - asks_t}{2} \]
correspond to the columns that correspond to the built indicators that were previously defined. For example, column is just a convenient variable that takes the same values as the closing prices. There are other Price* indicator the Indstdprc column describes the standard deviation price indicator describes the imbalance between bids and asks integrated the built the ask price, closing price and volume, as they were originally taken from the cited web site. Furthermore, we dataset.

We performed and ran the analysis using SPSS Clementine software. The analysis is based on the built financial indicators explained in the previous section; it mainly has three phases: No lags of touting email messages (Phase I), up to 5 days lag of the touting email messages (Phase II), and aggregate summary of the 5 days lag of the touting email messages (Phase III).

This work adopts the CRISP-DM model to define the problem of the spam scenario, selecting and understanding the dataset, preparing the dataset, modeling, and evaluating. Before performing any analysis the dataset requires data cleaning, integration and preprocessing. This kind of data preprocessing helps to produce quality data and a tuned dataset from the raw data before employing the model. As the most touting email messages occur from the period 02-01-2004 to 08-09-2005, we filtered the dataset to match this criterion and discard any records before the year 2004. This produces a table that contains the “CWTD” stock trading activities from 02-01-2004 to 08-09-2005. In addition, the built financial indicators have been added to the dataset.

Figure 3 represents the final filtered dataset which contains the date column, followed by the bid price, the ask price, closing price and volume, as they were originally taken from the cited web site. Furthermore, we integrated the touted column into the dataset to show the number of touted messages received in a particular day. Price* column is just a convenient variable that takes the same values as the closing prices. There are other columns that correspond to the built indicators that were previously defined. For example, Prc-Avg field for the second volume Indvol field for the second volume Indvolsd field, which calculates the standard deviation of the volumes in this stock.

\[
BA_{\text{var}2} = \begin{cases} 
1 & \text{if } BA_{\text{var}1} \geq 0 \\
0 & \text{if } BA_{\text{var}1} < 0 \\
null & \text{if } BA_{\text{var}1} = 0
\end{cases}
\]  

Finally, the Bids/Asks Indicator is calculated as:

\[
BA_{\text{Ind}1} = \frac{\sum_{i=1}^{18} BA_{\text{var}2} }{20}
\]

\[13\]

\[14\]

4 Analysis

We performed and ran the analysis using SPSS Clementine software. The analysis is based on the built financial indicators explained in the previous section; it mainly has three phases: No lags of touting email messages (Phase I), up to 5 days lag of the touting email messages (Phase II), and aggregate summary of the 5 days lag of the touting email messages (Phase III).

This work adopts the CRISP-DM model to define the problem of the spam scenario, selecting and understanding the dataset, preparing the dataset, modeling, and evaluating. Before performing any analysis the dataset requires data cleaning, integration and preprocessing. This kind of data preprocessing helps to produce quality data and a tuned dataset from the raw data before employing the model. As the most touting email messages occur from the period 02-01-2004 to 08-09-2005, we filtered the dataset to match this criterion and discard any records before the year 2004. This produces a table that contains the “CWTD” stock trading activities from 02-01-2004 to 08-09-2005. In addition, the built financial indicators have been added to the dataset.

Figure 3 represents the final filtered dataset which contains the date column, followed by the bid price, the ask price, closing price and volume, as they were originally taken from the cited web site. Furthermore, we integrated the touted column into the dataset to show the number of touted messages received in a particular day. Price* column is just a convenient variable that takes the same values as the closing prices. There are other columns that correspond to the built indicators that were previously defined. For example, Prc-Avg field for the second volume Indvol field for the second volume Indvolsd field, which calculates the standard deviation of the volumes in this stock.

\[
BA_{\text{var}2} = \begin{cases} 
1 & \text{if } BA_{\text{var}1} \geq 0 \\
0 & \text{if } BA_{\text{var}1} < 0 \\
null & \text{if } BA_{\text{var}1} = 0
\end{cases}
\]  

Finally, the Bids/Asks Indicator is calculated as:

\[
BA_{\text{Ind}1} = \frac{\sum_{i=1}^{18} BA_{\text{var}2} }{20}
\]

\[13\]

\[14\]
The data mining techniques that have been utilized in this study required the existence of a categorical field to functionally work. In this case, we set a new field to the dataset called Toutlabel based on a conditional format. This condition classifies the touted records into four categories high \((touted\geq30)\), medium \((15\leq touted<30)\), low \((0< touted<15)\), and notout \((touted=0)\). We set a conditional format based on a histogram for the touting email messages as shown in figure 4. The cutting point was 50 email messages.

![Figure 4. Toutlabel Histogram](image)

Logically, the touting should not occur very often; thus, this massive amount of notout records should be balanced. As shown in the previous histogram the green bar describes the notout category which has the highest proportion in the dataset. Unfortunately, this may have a negative impact and misdirect the techniques to predict the highest possibilities of notout transactions. For that reason, we adjust the current balancing directives of the dataset and decrease the proportion of notout records in the training dataset. It is minimized to 25% of the records and set the proportion of High and Medium to use the whole sample (100%) which is the study objective.

In our study, we used the lag analysis to examine the effect of thetout spam on the “CWTD” stock after a few days of dissemination. The objective of this analysis is based on our assumption that tout spam could heavily affect the trading behavior of the stock in the days following the email distribution. Therefore, we started our analysis by examining the impact of tout spam on the cited stock at the same day of the email distribution. Additionally, we evaluated the affect of stock spam on “CWTD” through five consecutive trading days. This helps to forecast the moving average of email touting which pinpoints any changes in the stock behavior. Lastly, this study employed the aggregation of the 5 days lag as another analysis to evaluate the abnormal behaviours of the spam data in the whole 5 days at the same time.

The experiment can be described as follows: Given the fact that “CWTD” stock has been touted form 02-01-2004 to 08-09-2005, we employ data mining techniques which should be able to predict the highest possibilities of High and Medium categories in the “toutlabel” field of this stock. For the validity and reliability of study, we train 50% of the dataset as a random sample to evaluate the whole analysis and eventually generate a model then test it. This is a real world scenario where the data preparation has been revisited more than once for the amendments needed to fit the modeling approach. Furthermore, it will help in training the model on a sample period then predicting the subsequent time period.

The algorithms were evaluated: C5.0 for building a decision trees, neural network, and logistic regression. All models were built using SPSS Clementine as mentioned before. C5.0 model was run with the default parameters. The simple training mode with accuracy favour option was selected. Furthermore, C5.0 parameters were set automatically using pruning severity of 75% and minimum of 2 records per child branch to produce the most accurate possible fraudulent tree. The “toutlabel” field decision tree is the result of the generated model. Secondly, neural network were used with the “prune” method. This method begins with a large network and removes (prunes) the weakest units in the hidden and input layers as training proceeds. In addition, the three hidden layer were utilized in neural network prediction. This study used 100% of the specified proportion as a sample of the data to train the neural network. However, this method is often slow, but it usually gets better results than other methods. Lastly, logistic regression was used with the “multinomial” procedure because “toutlabel” field is a set field that has more than two values. Furthermore, the method that has been used in this model is the “backwards” method that initially contains all of the terms as predictors, and terms can only be removed from the model. Basically, it initiates a model contains all features and iteratively removes feature that do not significantly add value to the model. Again, the default Clementine parameters were employed.
5  Discussion and Results

This study evaluates and discusses the ability of the generated model to engender accurate predictions based on the aforementioned analysis. Figure 5 demonstrates the results for the output field “toutlabel”. The top level section contains a table showing the number and percentage of correct and incorrect rate predictions and the total number of records that have been processed. Furthermore, within the “toutlabel” field section is a subsection for each prediction field associated with that output field. Give the fact that “toutlabel” field is a categorical output, the predictions rates were based on scoring records and comparing the responses predicted by the model to the actual results of the “toutlabel” field. This contributes to evaluate the generated model as will be explained in the following sections.

We began our analysis by examining the abovementioned data mining techniques through three phases (phase I, phase II, phase III) before partitioning the dataset. The results are summarized in table 1. Phase I experiment shows that C5.0 technique identified 14 (3 high, 11 medium) abnormal behaviours over the full range of data. In this analysis C5.0 scored 58% correct prediction rate. Neural network techniques were able to detect 19 cases with a correct rate of 66%. Logistic regression was slightly worse, detecting only 2 high touting labels with 46% correct rate.

Phase II experiment shows that C5.0 successfully detected fraudulent cases from the whole sample with a correct rate ranged from 66% to 71%. For example, at the first day of outing (1st day lag) after the email distribution C5.0 could recognized 27 high and medium touted labels. Second day (2nd day lag) shows a slight decrease to 25 touted stocks. The number of outing decreased to 15 in the third day and rose up again to 26 touted in the fourth day. On the fifth day it decreased again to detect only 18 cases. In contrast, the neural network did not come up with any touted stock for the whole 5 days with a correct rate ranged from 44% to 52%. At the same time the logistic regression was just able to detect 2 touted stocks at the first day and only one outing in the second day with correct rate ranged from 41% to 48%.

Phase III shows that C5.0 generated impressive results in this stage by detecting 47 touting email messages with a correct rate of 71%, which is the best detection result in the whole analysis. Neural networks were able to detect 13 cases with a correct rate of 50%. Unfortunately, logistic regression only picked one case with a correct rate of 45%.

Table 1. The Summary of Primary Results

<table>
<thead>
<tr>
<th>Analysis Stages</th>
<th>C5.0</th>
<th>Neural Network</th>
<th>Logistic Regression</th>
<th>C5.0 Correct rate</th>
<th>Neural No Correct Rate</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>No days Lag</td>
<td>14 (3 High, 11 medium)</td>
<td>13 (3 High, 16 medium)</td>
<td>2 (2 High)</td>
<td>50%</td>
<td>66%</td>
<td>46%</td>
</tr>
<tr>
<td>1st day lag</td>
<td>27 (11 High, 16 medium)</td>
<td>0</td>
<td>2 (2 High)</td>
<td>71%</td>
<td>47%</td>
<td>48%</td>
</tr>
<tr>
<td>2nd day Lag</td>
<td>25 (5 High, 20 medium)</td>
<td>0</td>
<td>1 (1 High)</td>
<td>69%</td>
<td>51%</td>
<td>47%</td>
</tr>
<tr>
<td>3rd day Lag</td>
<td>15 (7 High, 8 medium)</td>
<td>0</td>
<td>0</td>
<td>63%</td>
<td>52%</td>
<td>55%</td>
</tr>
<tr>
<td>4th day Lag</td>
<td>26 (9 High, 17 Medium)</td>
<td>0</td>
<td>0</td>
<td>72%</td>
<td>52%</td>
<td>48%</td>
</tr>
<tr>
<td>5th day Lag</td>
<td>18 (6 High, 12 Medium)</td>
<td>0</td>
<td>0</td>
<td>66%</td>
<td>44%</td>
<td>42%</td>
</tr>
<tr>
<td>Aggregate of 5 days</td>
<td>87 (17 High, 10 medium)</td>
<td>13 (0 High, 15 medium)</td>
<td>1 (1 Medium)</td>
<td>71%</td>
<td>30%</td>
<td>45%</td>
</tr>
</tbody>
</table>

Table 2 shows the summary of the training analysis results (50% of the sample) using the three data mining techniques. Overall, the three techniques in phase I have not performed well in this analysis. C5.0
technique detected only 2 high touted labels as abnormal behaviours of the “CWTD” stock with 58% correct rate. Neural network were not able to pick any cases of high or medium touting with a correct rate that reached 43%. Logistic regression detected just one high touting label with a total of 48% correct rate.

However, in phase II C5.0 successfully detected abnormal behaviours in this stage with a correct rate ranged from 64% to 69%. For the first day of touting the C5.0 came up with 11 touted labels. The second day shows an increase to 25 touted stocks. The number of touting decreased reaching 15 in the third day and no cases were detected in the fourth day. For the fifth day the number of touting reached 13 cases. Unfortunately, the neural networks were able to detect 4 cases only in the first day and nothing in the other days, with a correct rate range from 44% to 52%. At the same time, the logistic regression was able to detect 3 touted stocks in the first day and 2 touting in the second day, failing to detect any other cases for the rest of the days with a correct rate ranged from 45% to 57%.

In phase III, C5.0 shows a significant success in this stage by detecting 24 touting email messages with a correct rate of 71%. At the same time, neither the neural networks nor logistic regression were able to detect any abnormal case.

Table 2. The Summary of Training Analysis Results

<table>
<thead>
<tr>
<th>Analysis Stages</th>
<th>C 5.0</th>
<th>Neural Network</th>
<th>Logistic Regression</th>
<th>C 5.0 Correct Rate</th>
<th>Neural Nw Correct Rate</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>No days Lag</td>
<td>2 (2 high)</td>
<td>0</td>
<td>1 (1 High)</td>
<td>58%</td>
<td>43%</td>
<td>48%</td>
</tr>
<tr>
<td>1st day Lag</td>
<td>11 (5 High, 6 medium)</td>
<td>4 (4 High)</td>
<td>3 (2 High, 1 Medium)</td>
<td>64%</td>
<td>53%</td>
<td>57%</td>
</tr>
<tr>
<td>2nd day Lag</td>
<td>12 (2 High, 10 medium)</td>
<td>0</td>
<td>2 (2 High)</td>
<td>66%</td>
<td>51%</td>
<td>48%</td>
</tr>
<tr>
<td>3rd day Lag</td>
<td>10 (4 High, 6 medium)</td>
<td>0</td>
<td>0</td>
<td>69%</td>
<td>48%</td>
<td>48%</td>
</tr>
<tr>
<td>4th day Lag</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>55%</td>
<td>42%</td>
<td>40%</td>
</tr>
<tr>
<td>5th day Lag</td>
<td>13 (3 High, 10 Medium)</td>
<td>0</td>
<td>0</td>
<td>69%</td>
<td>46%</td>
<td>45%</td>
</tr>
<tr>
<td>Aggregate of 5 days</td>
<td>24 (6 High, 18 medium)</td>
<td>0</td>
<td>0</td>
<td>71%</td>
<td>46%</td>
<td>54%</td>
</tr>
</tbody>
</table>

Table 3 demonstrates the results of the testing analysis (the other 50% of the sample). In phase I, C5.0 technique detected 8 touted labels as abnormal behaviours of the “CWTD” stock with 60% correct rate. At the same time, neural networks were not able to get any cases with a correct rate of 47%. However, Logistic regression was able to detect three touting label with 52% correct rate.

It is clearly seen in phase II that C5.0 was successful in detecting abnormal behaviours for this sample, with a correct rate ranged from 64% to 68%. For the first day touting, C5.0 picked nine cases. The second day shows a dramatic decrease reached to two cases. The number of touting rose up to reach eleven cases in the third day and continue increasing up to fifteen cases in the fourth day. The fifth day reached eleven touting stocks. Unfortunately, the neural networks were unable to detect any cases in this stage. At the same time, the logistic regression was able to detect two touted stocks on the fourth day and three in the fifth day with correct rate ranged from 43% to 52%.

Again in phase III, C5.0 performed well by detecting twenty eight touting email messages with a correct rate reaching 83%. Neural networks did not detect any abnormal case in this stage. Additionally, logistic regression picked five cases only with 46 % correction rate.

Table 3. The Summary of Testing Analysis Results

<table>
<thead>
<tr>
<th>Analysis Stages</th>
<th>C 5.0</th>
<th>Neural Network</th>
<th>Logistic Regression</th>
<th>C 5.0 Correct Rate</th>
<th>Neural Nw Correct Rate</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>No days Lag</td>
<td>8 (2 high, 6 medium)</td>
<td>0</td>
<td>3 (1 High, 2 Medium)</td>
<td>60%</td>
<td>47%</td>
<td>52%</td>
</tr>
<tr>
<td>1st day Lag</td>
<td>9 (6 High, 3 medium)</td>
<td>0</td>
<td>0</td>
<td>60%</td>
<td>46%</td>
<td>46%</td>
</tr>
<tr>
<td>2nd day Lag</td>
<td>2 (2 High)</td>
<td>0</td>
<td>0</td>
<td>60%</td>
<td>42%</td>
<td>45%</td>
</tr>
<tr>
<td>3rd day Lag</td>
<td>11 (3 High, 8 medium)</td>
<td>0</td>
<td>0</td>
<td>60%</td>
<td>41%</td>
<td>43%</td>
</tr>
<tr>
<td>4th day Lag</td>
<td>15 (6 High, 9 medium)</td>
<td>0</td>
<td>2 (1 High, 1 Medium)</td>
<td>65%</td>
<td>46%</td>
<td>56%</td>
</tr>
<tr>
<td>5th day Lag</td>
<td>11 (6 High, 7 Medium)</td>
<td>0</td>
<td>3 (2 High, 1 Medium)</td>
<td>64%</td>
<td>46%</td>
<td>52%</td>
</tr>
<tr>
<td>Aggregate of 5 days</td>
<td>28 (11 High, 17 medium)</td>
<td>0</td>
<td>5 (3 High, 2 Medium)</td>
<td>83%</td>
<td>46%</td>
<td>46%</td>
</tr>
</tbody>
</table>

As it was expected, some techniques performed better than others for this particular problem. Specifically, the C5.0 technique outperformed the alternatives techniques in the phases of the analysis. Neural networks and linear regressions were slightly worse, but they still may be suitable for other types of problems. These results are in line with previous results, such as the research of [2], in terms of the order and the power of classification the techniques. As shown in [2], the C4.5 algorithm also outperformed the neural networks, and the logistic regression, for a problem of classification for early detection of insider trading in the option market. At this stage of the study, it is premature to confirm which one of the five proposed financial indicators was the best.
for the modelling phase. All indicators performed well in the analysis and helped the techniques for the classification task.

The results presented in this section suggest that some investors actually do respond to the spam stock touts, in concordance with previous spam research [3] and with the accumulated effect captured in the 5 days lag analysis. The experiment shows that aggregate summary of the 5 days lag analysis has the highest correct rates range from 71% to 83% of the C5.0 technique, as well as a successful analysis to detect the highest possibilities of abnormal behaviour in all phases of the spam case as highlighted in the tables 1, 2 and 3. Regarding the reliability and the validity of the analysis, especially the testing phase results prove that the data mining models and techniques can work and detect fraud patterns of this kind of problem.

6 Conclusions

Spammers are utilizing the Internet as a framework to target large number of investors. Spam emails have been used as one of the information-based manipulation tool. Spammers (manipulator) are working to drive the price up or down and change the value of the announced stock and gain illegal profit. Therefore, this study uses the China World Trade Corporation “CWTD” stock as a case study to detect the highest possibilities of fraudulent activities. This paper focus on utilizing data mining techniques to detect any fraudulent activities.

The experiment shows that the proposed analysis is able to identify the highest possibilities of High and Medium categories in the “toutlabel” fields of the “CWTD” stock. The results strongly suggest that the data mining models and techniques can work and detect fraud patterns for these kinds of problem. Particularly, the analysis shows that C5.0 technique outperformed the alternatives techniques in all the phases. Neural networks and linear regressions were slightly worse, but they still may be suitable for other types of problems. The built financial indicators performed well and helped the techniques to work properly and produced a reasonable results.

The analysis demonstrates that aggregate summary of the 5 days lag analysis has the top correct rates range from 71% to 83% of the C5.0 technique, as well as a successful analysis to detect the highest possibilities of abnormal behaviour in the two phases of the spam case. That could be an indicator of some investors acting upon the spam stock touts, in the same line with previous spam research of [3] and with the accumulated effect captured in the 5 days lag analysis.

Following the work reported in this paper, there are a number of issues that need to be addressed through further investigation. In particular, the research did not have access to neither the intra-day price transactions nor the number of bid or ask offers of “CWTD” stock that can provide details about any suspicious price movements occurring during the day. Additionally, the date range did not allow the research to look for re-occurring instances of trading anomalies. The paper did not consider the control of the news effects. The research focused only on one market (Pink Sheets market) which is characterized as an unstructured market. Furthermore, the techniques and the algorithms used in this research can be expanded to include other techniques such as neural networks, kappa statics and others.

Finally, further research could address additional financial indicators based on the intra-day price transactions. Future work could also consider using text mining techniques to analyze the common characteristics of the spam emails. The combination of different sophisticated techniques and algorithms is expected to increase the accuracy of detecting different manipulation patterns.

References


