Algorithmic Finance 2 (2013) 169–196
DOI 10.3233/AF-13025
IOS Press

Stock chatter: Using stock sentiment to predict price direction

Michael Rechenthin*[a], W. Nick Street[a], and Padmini Srinivasan[b]
[a]*Department of Management Science, The University of Iowa, 108 John Pappajohn Business Building, Iowa City, IA, USA, Email: mrechenthin@gmail.com*
[b]*Department of Computer Science, The University of Iowa, 14 MacLean Hall, Iowa City, IA, USA*

Abstract. This paper examines a popular stock message board and finds slight daily predictability using supervised learning algorithms when combining daily sentiment with historical price information. Additionally, with the profit potential in trading stocks, it is of no surprise that a number of popular financial websites are attempting to capture investor sentiment by providing an aggregate of this negative and positive online emotion. We question if the existence of dishonest posters are capitalizing on the popularity of the boards by writing sentiment in line with their trading goals as a means of influencing others, and therefore undermining the purpose of the boards. We exclude these posters to determine if predictability increases, but find no discernible difference.

Keywords: prediction, classification, sentiment analysis, stock, equity, tweet, Yahoo finance, message boards

1. Introduction

Stock message boards give users a place to ask questions, find useful information and discuss rumors regarding a chosen stock. Of the stock message boards on the internet, Yahoo hosts one of the largest and most popular online communities, with boards for over 6000 stocks. Participation on the Yahoo boards varies from posts once a month to thousands of posts per day. Posts tend to be shorter and less thoughtfully written on average than a newspaper article, and the discussions tend to be more conversational in nature. In addition to contributing written posts, the poster on the Yahoo message boards has the ability to choose a long-term sentiment disclosure of his/her recommendation for readers (i.e. "strong buy", "buy", "hold", "sell", and "strong sell"). This sentiment is displayed at the bottom of the message (see Figure 1) and is referred throughout this paper as "explicit sentiment."

There are numerous reasons why individuals participate in conversations on message boards. The primary reason is to exchange ideas regarding publicly traded stocks. Das and Chen (2007) mention that while large institutions express their opinions on a stock via published stock forecast, stock message boards provide smaller investors a place to converse and share their opinions. Cao et al. (2002) argue that participants of the market do not always have the confidence to act upon their trading ideas. These sidelined individuals seek out others who share similar opinions to gain confidence to trade. Message boards may therefore be the ideal venue for these individuals to interact. It is this interaction that may play a role in the movement of stock prices. As we know from basic psychology, emotion plays a significant role in the decision making process. A message board post or news article may influence an investor's or trader's emotion which may indirectly influence the stock's price. Antweiler and Frank (2004) and Das and Chen (2007) find evidence that seems to support this hypothesis that message board activity coincides with stock volatility[1].

*mrechenthin@gmail.com

[1]Volatility describes the variability (i.e. the standard deviation) of the stock returns, or the magnitude and speed of the stock price fluctuations. Das and Chen calculate volatility as the difference between the high and low stock prices for the day divided by the average of the open and closing price. Antweiler and Frank use a measure of volatility, called the Sharpe ratio, which is a measure of return per unit of deviation in excessive of a benchmark. Both measures of volatility are popular.

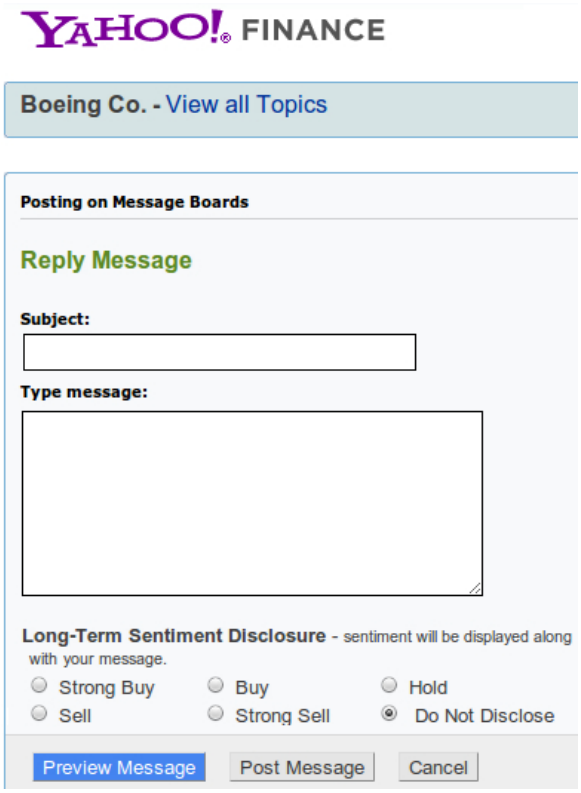


Fig. 1. Example of posting to the message board. Explicit sentiment can be decided by the poster; the default is "Do Not Disclose."

It is one thing to predict volatility, yet another and far more interesting, to predict price direction. Recent research examining crowd-sourced information contained within blogs and message boards to predict the direction of stock market prices has shown positive results. Schumaker and Chen (2009) used linguistic and statistical techniques to predict market direction using news articles and produced a small profit using their model when tested against a benchmark. Choudhury et al. (2008) using a SVM showed that the activity on blogs are correlated with the underlying stock direction. Li et al. (2011) uses a multi-kernel learning algorithm on news articles and historical price information to improve a baseline model. Wex et al. (2013) used 45 million Reuters news stories over eight years to help explain oil price movements. Additionally, Balakrishnan et al. (2010) used quarterly financial filings to predict a stock's performance. The sentiment and tone of the filings were found to be slightly predictive of future stock returns.

Other than traditional blogs and message boards, Twitter has recently become a popular source of data for finding market predictability. For example in Ruiz et al. (2012) the authors found high levels of correlation between Twitter and stock volumes and this along with the stock's price was shown to produce a small profit. According to (Brown 2012; Ruiz et al. 2012; Sprenger & Welpe, 2010) the loosely based convention on Twitter when discussing stocks is to use a hashtag followed by the stock's symbol (i.e. #AAPL, $AAPL, or #Apple). However, the lack of consistency among posters in hashtag use, and the high level of noise makes finding post related to a specific stock difficult.

An example of this difficulty in finding tweets relating to a specific stock is Apollo Group Incorporated (symbol: APOL). A Twitter search finds post related to Apollo Theater in Harlem, Apollo the Greek god, Apollo 13 manned mission to the Moon, and numerous unrelated companies. In (Ruiz et al., 2012) the authors used different hashtags, along with several filters, such as if the tweet uses the word "stock", to find relevant, stock specific posts, but also states "when we determined that a rule-based approach was not feasible, we removed that company from our dataset." This difficulty prevented the use of Twitter as a source in our research.

(a) MarketWatch.com (b) SentimenTrader.com

Fig. 2. Sentiment indications being displayed for the market on popular websites for traders.

Many papers have instead examined Tweets not related to a particular stock, but instead to the overall stock market. For example, Bollen et al. (2011) examined ten million Twitter feeds to predict the overall market movement up to six days in advance using six dimensions of mood. Zhang et al. (2011, 2012) also used Tweets confined to the United States to find overall slight market predictability.

With the potential for huge profit, it is no surprise that hedge funds and large institutional traders are looking into the use of sentiment analysis within their trading models. According to Aite Group, a financial services consulting company, as of the end of 2010, 35% of professional trading firms were exploring the use of sentiment analysis in their models, up from 2% in 2008 (Bowley, 2010). Several stock trading websites display a crowd-sourced sentiment that is displayed when a user inputs a stock (see Figure 2 for two sentiment detecting websites). Examples of websites that provide sentiment analysis for stocks (either for free or for a subscription fee service) include: DataMinr, Bloomberg, MarketWatch, The Motley Fool, and The Stock Sonar[2]. There is, however, a lack of transparency on how these websites calculate positive versus negative stock sentiment. With the incentive being profit, a dishonest message board poster could undermine the system to make stocks appear to have sentiment inline with their trading goal ("bullish" sentiment if they *own* the stock or "bearish" sentiment if they are *short* the stock and want to push the stock down) as a means of influencing others.

Message boards can be abused, since users can post anonymously. DeMarzo et al. (2003) argue that people give more weight to the opinions of those with whom they talk and this kind of belief makes it profitable to be an influential participant within the message boards. Short sellers (those who profit when the stock drops in price) trying to frighten others into panic selling are mixed into the boards. Arther Levitt, former Security and Exchange Commission (SEC) Chairman stated "I encourage investors to take what they see over chat rooms not with a grain of salt but with a rock of salt." An act where individuals disseminate false information through message boards and/or email and then sell their stocks at artificially inflated prices is a "pump and dump" scam (Gu, Konana, Liu, Rajagopalan, & Ghosh, 2006). By enticing others to buy the stock, the scammers create a high demand for the stock which raises the price. This sudden increase in the stock's price entices others to believe the hype and to buy shares as well. When the individuals behind the scheme sell their shares at a profit and stop promoting the stock, the price plummets, and other investors are left holding stocks which are worth significantly less than what they paid for it. A study by Frieder and Zittrain (2007) examined stocks where "pumpers" previously sent large quantities of emails to entice others to buy, and found that investors who bought the stocks lost,

[2]DataMinr, www.dataminr.com, recently raised $30 million and partnered with Twitter to identify credible tweets; Bloomberg, www.bloomberg.com, is one of the largest financial data services; MarketWatch, www.marketwatch.com, is owned by Dow Jones & Company; The Motley Fool, www.fool.com, since 1993 has been providing financial advice to individual investors and has a syndicated newspaper column along with two New York Times best sellers; and The Stock Sonar, www.thestocksonar.com, is a popular sentiment website and also mentioned in (Feldman, Rosenfeld, Bar-Haim, & Fresko, 2011).

on average, 5.25% in the two day period following touting.

A famous case of "pump and dump" involved 15-year old high school student Jonathan Lebed who would purchase stocks in advance, and then send spammed touts on Yahoo message boards on the same day. His six-month trading profits amounted to $800,000. The SEC settled a case against Jonathan in which he had to give up a portion of his profits, with interest (Lewis, 2001). This profit potential in stocks for the individuals doing the touting, creates an ideal environment for abuse and demonstrates the importance to filter spammers and other irrelevant posts.

The main contribution of this paper is to examine posts (and their posters) of eleven popular stocks on the Yahoo Finance message boards. The value of the message boards itself can be established by the percentage of on-topic communication and the level of underlying stock predictability, as determined by the sentiment contained with the posts. Using supervised learning algorithms, we use the sentiment contained in the message board to attempt to predict the underlying stock price and find slight predictability when combined with historical price information. Lastly, using four metrics of finding outliers among posters, we find posters whose post raise suspicion. We theorize this may be the existence of possible "pump and dumpers" (participants whose intent is to influence others to artificially inflate stocks prices). These posters are excluded to determine if predictability increases, but we find no discernible difference.

The rest of the paper is organized as follows: In Section 2 we empirically examine nearly 70,000 Yahoo Finance message board posts over eleven stocks and their near 7,000 posters to determine if the data can aid in the discovery of worthwhile knowledge. In Section 3 we discuss the use of supervised learning methods to determine the sentiment of posts and compare this with the posters own explicitly provided sentiment. In Section 4 the best performing sentiment model is used to determine if predicted and/or explicit sentiment can be used to predict the future market price. Section 5 examines users who we surmise to be possible "suspicious" users due to four metrics. The first is the frequency of posts; a user who posts over a certain criterion is deemed an outlier. The second metric is username similarity. We found users who had high Levenshtein username similarity to others wrote a statistically significant greater number of posts and provided more sentiment. Third, we examined text similarity of posts by different users. Less than seven percent of users had high levels of similarity but these users wrote over 52% of messages. The last metric we examined for finding potential suspicious users was the length of time the account was opened. With the ease of opening a Yahoo account, we expected greater number of spam messages from short-term accounts; however, this is not what was found. Lastly we exclude the "suspicious" users to determine if predictability increases but find no discernible difference.

2. Overview of boards

2.1. Data collected

We examined $67,849$ posts regarding eleven widely-traded stocks over a varying timespan, which can be seen in Table 1. This represented the entire collection of messages posted during the timespan for the particular stock. The stocks were chosen due to their large number of posts and high trading volumes. The average number of posts among stocks was $145[\pm 98]$ per day, with the least being Bristol-Myers Squibb with an average of $40[\pm 23]$ per day and Cisco being the highest with an average of $445[\pm 228]$ per day. The posts were retrieved using a program written in Java by one of the authors that utilized the open-source JSoup API. In addition to storing the user's posts, the message title, the date and time of posts, and the users explicit sentiment (if provided), we retrieved user specific information such as the date their Yahoo account was opened.

It is important to note that Yahoo does monitor its message boards along with the help of its members. A "report abuse" link on posts allows users to report messages that go against Yahoo's terms-of-service

Table 1
Stocks examined for this study

Stock	Time frame	Posts	Post/day
Boeing (BA)	April 23 - July 28 2011	9249	$95[\pm 77]$
Bristol-Myers Squibb (BMY)	April 26 - July 28, 2011	3796	$40[\pm 23]$
Cisco (CSCO)	May 25 - June 4, 2011	4891	$445[\pm 228]$
Google (GOOG)	April 15 - June 1, 2011	4180	$87[\pm 119]$
Intel (INTC)	May 4 - May 31, 2011	4435	$158[\pm 117]$
Microsoft (MSFT)	May 16 - June 4, 2011	3553	$178[\pm 72]$
Nokia (NOK)	April 21 - July 28, 2011	13742	$139[\pm 170]$
Pfizer (PFE)	April 11 - May 31, 2011	4481	$88[\pm 57]$
Wal-Mart (WMT)	May 1 - July 28, 2011	12762	$88[\pm 57]$
Exxon (XOM)	April 27 - June 1, 2011	2326	$65[\pm 32]$
Yahoo (YHOO)	May 12 - June 3, 2011	4504	$196[\pm 169]$

agreement; this includes high levels of profanity, racism, threatening language or other forms of abuse. In addition, messages that include promotion of a commercial website (spamming), libel, off-topic comments, and the publishing of personal information that puts others at risk of fraud is subject to deletion. Nevertheless, we take a random sample of $1,677$ messages and find that over 68.8% of those messages escape filtering and are classified by us as off-topic. We will discuss this further in the paper.

2.2. Posts

2.2.1. Typical post

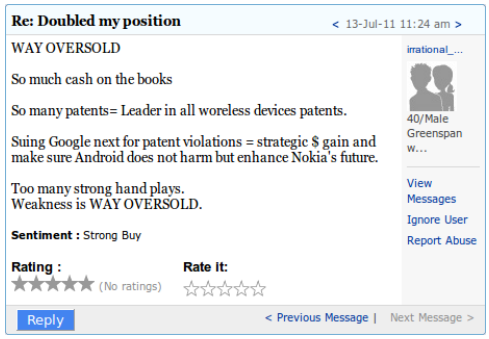
Consider the following post by user, "irrational_fed_selloff", posted on July 13, 2011 on the Nokia (stock symbol: NOK) message board (see the message also in Figure 3a):

> WAY OVERSOLD So much cash on the books. So many patents = Leader in all woreless devices patents. Suing Google next for patent violations = strategic $ gain and make sure Android does not harm but enhance Nokia's future. Too many strong hand plays. Weakness is WAY OVERSOLD.

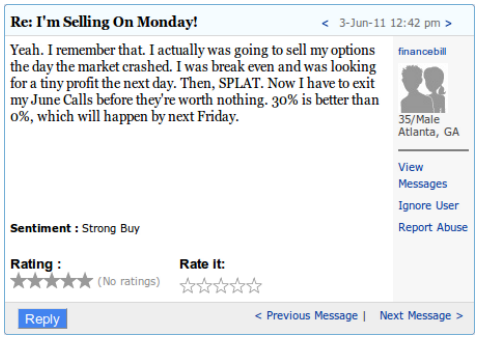
Notice the poor grammar, preference for capital letters, and misspellings within the message; this is quite common in the examined posts. In addition to containing an explicit sentiment of "strong buy" which was provided by the user, the context of the post would lead most to interpret this as a positive message. First, the poster mentions that the company is "way oversold" and second, he provides statements that would be interpreted by most as positive – strong cash position, worldwide leader, and a strategic goal of reducing competition through legal maneuvers (i.e. suing Google). A second example is given in Figure 3b. However in this post, the author is quite negative toward the stock, yet selects a "strong buy" explicit sentiment. This is either a mistake (i.e. he clicked the wrong box), or the user is attempting to mislead sentiment-capturing websites that use a more naïve algorithm of capturing sentiment only as a count of the explicit sentiment. Later in this paper we use evaluators to manually label the implied sentiment provided by the user within the text of the post to determine how often this happens.

Posts on average included $42[\pm 72]$ words with a median of 19 words and 90% of the posts under 100 words (see Figure 4). There are a total of $60,326$ unique words in the collection of $67,849$ post. Replies to existing posts on the message boards made up $65.1\%[\pm 9.4\%]$ of the total number of posts on the board, with an average of $3[\pm 4]$ replies per posts. As pointed out by Antweiler and Frank (2004), it is possible that the structure of the Yahoo Finance message boards makes it easier for users to "reply" to earlier messages rather than create his/her own messages. Only on the main message board page can users "Start a new topic", whereas users can "Reply" on any page.

The high number of replies to existing messages creates a complex multi-threaded structure of communication within the posts and often within the same topic header. As explained in Wang et al. (2008), if Participant C disagrees with B's disagreement with A's opinion, then C agrees with A. An individual's reply could be could be seen as in agreement toward another post in the thread that is negative toward the subject. This adds complexity to the problem.



(a) Implied positive sentiment and explicit positive sentiment ("strong buy")



(b) Implied negative sentiment, but explicit positive sentiment ("strong buy")

Fig. 3. Examples of implied and explicit sentiment given within posts.

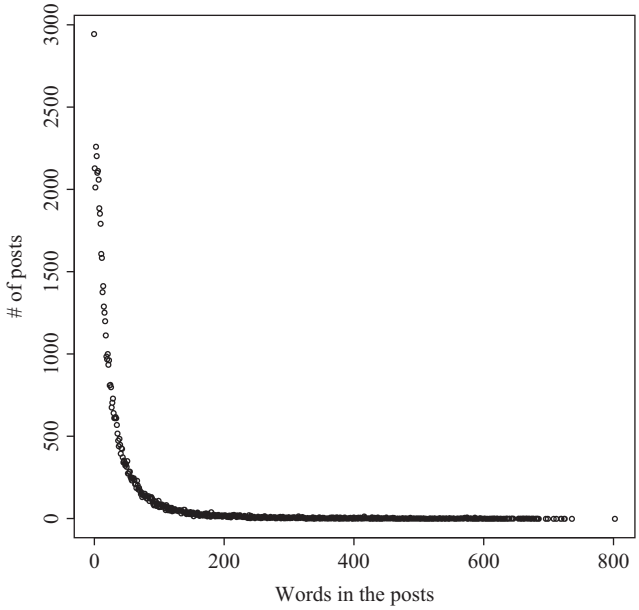


Fig. 4. Word count among posts.

Furthermore, $10.6\%[\pm 4.1\%]$ of all posts including links to other articles or websites. Links make analysis more difficult, since a true analysis of the posts would also follow the link to which the article is referring. In this paper, we treat all posts as independent of one another within the algorithm, but did ask that the individuals who helped us classify the training set posts take into consideration the thread in which the posts belonged.

To understand how the words are distributed across the message board posts, Zipf's law is a commonly used model. This is the empirical principal where the j^{th} most common word is proportional to $1/j$. The 10 most common words in the collection are *the, to, and, a, of, in, is, that, you, it.* Figure 5 displays $\log_{10}$ rank on x-axis and $\log_{10}$ collection frequency on the y-axis. While the terms in the message board posts do not fit the distribution perfectly, it does well enough to provide for approximations.

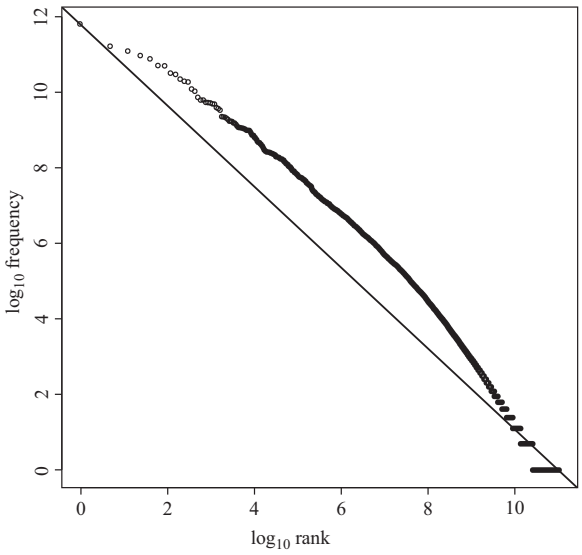


Fig. 5. Examining Zipf's law for the collection.

2.2.2. Daily and hourly distribution of posts

As can be expected, the majority of posts are on market trading days (Monday through Friday) with 84.5% of posts (see Figure 6). The ending of the trading week (Thursday and Friday) appears to have more activity, but this is not statistically significant. Regular trading activity on the New York Stock Exchange is 08:30 to 15:00 hours Central Standard Time. From Figure 7 it can be observed that over half ($61.2\%[\pm 9.9\%]$) of all messages that are posted during the week are posted during opening hours, with a tapering-off of messages after the market closes. Similar results were found by Antweiler and Frank (2004) where they theorize these messages could be due to day traders or individuals checking and conversing about their stocks from their work. In addition, messages posted during the evening may be from smaller and less active traders who are posting when they are off work. The high level of

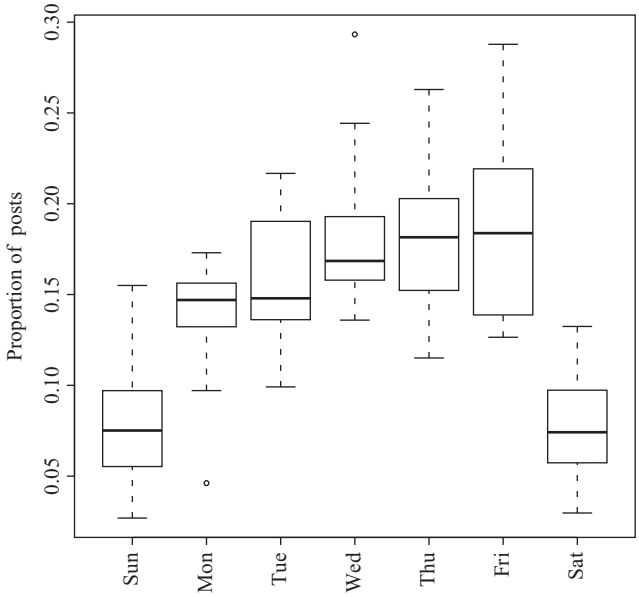


Fig. 6. Boxplot of stocks showing the distribution of daily message board activity for the 11 stocks.

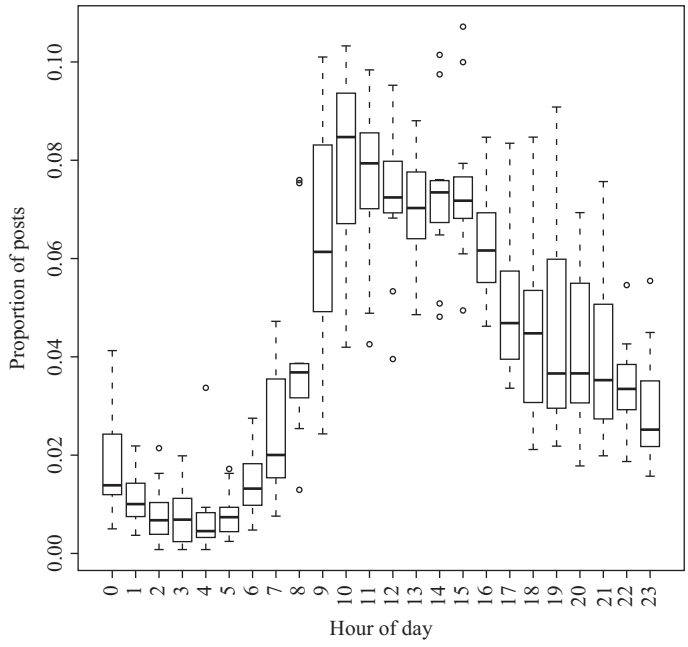


Fig. 7. Boxplot of stocks showing the distribution of hourly message board activity for the 11 stocks.

off-topic conversations on the message board outside of market hours, which we will later discuss, gives credibility to this theory.

2.2.3. Posters

Varying levels of activity among the posters are found in the boards. 51.1% of posters write one message during the course of the dataset timespan, which as a proportion, accounts for a total of 5.2% of the posts. 1.8% of users write 100+ messages, which account for a total proportion of 47.3% of posts. The top poster to the individual boards alone writes from 7.01% to 21.49% of all the posts for the observed stocks during the observed timespans. In addition,

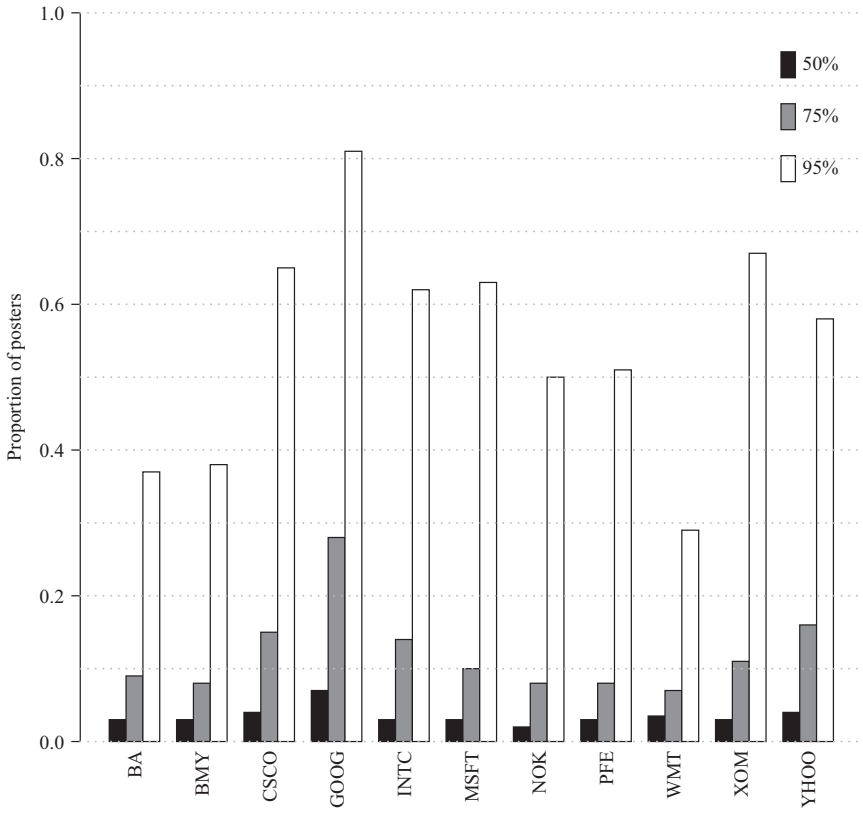


Fig. 8. The % of messages explained by the proportion of posters broken down by stock symbol.

50% of posts are written by just 2.9%[±1.3%] of the posters, 75% of posts by 11.2%[±5.8%] of posters, and 95% of posts are written by 54.2%[±15.7%] of posters.

In Figure 8, we show the percentage of messages by proportion of posters broken down with stock symbol. From this chart, we can observe that in the Wal-Mart (symbol: WMT) message board, 95% of all the messages are written by 29% of posters, while in the stock Google (symbol: GOOG), 95% of the messages are written by a much larger proportion, 81%, of posters. If the board is truly open and active by many posters with many different ideas, then it would be expected to have a large percentage of posters writing the messages such as Google (Symbol: GOOG). On the other end of the spectrum, a message board that is mostly dominated by a few would have a large percentage of posts written by a small minority of posters such as Wal-Mart (Symbol: WMT).

2.2.4. Explicit and implied sentiment consistency

As mentioned previously, posters can explicitly choose a sentiment along with their posts. These sentiment options are "strong buy", "buy", "hold", "sell", or "strong sell" (see Table 2). On average 17.3%[±8.4%] of the observed stocks' posts included explicit sentiment, with bullish explicit sentiment ("strong buy" or "buy") being most common in 7 out of the 11 stocks (see Table 3). Boeing (stock symbol: BA) had the lowest level of users supplying explicit sentiment with their message, with Intel (stock symbol: INTC) posters supplying the highest as a percentage of messages posted. Extreme sentiment, such as "strong buy" and "strong sell" are more common when the user has explicitly provided a sentiment, than are sentiments of "hold", "buy" and "sell." This can be seen in Figure 9.

The question remains: What is the sentiment of the remaining 82.7% of the posts that do not include explicit sentiment, and furthermore, is the author-provided explicit sentiment representative of the poster's writings within the posts? How often do users post explicit sentiment of "strong buy", yet write a bearish sentiment, as seen in Figure 3b? Answering these question can help us determine the suitability of the message boards in predicting stock prices.

Table 2
Different sentiments that can be explicitly attached to the Yahoo message posts by the poster

Posters Explicit Sentiment	Meaning
Strong buy	Poster has a bullish or positive
Buy	outlook on the stock
Hold	Poster has a neutral outlook on the stock
Sell	Poster has a bearish or negative
Strong sell	outlook on the stock
Not included	Poster does not include sentiment on the stock

Table 3
Distribution of explicit sentiment given per stock

Symbol	Strong buy	Buy	Hold	Sell	Strong sell	Not included
BA	0.9%	0.5%	0.1%	0.1%	3.0%	95.3%
BMY	1.6%	1.9%	0.2%	5.4%	0.2%	90.6%
CSCO	10.7%	1.2%	1.2%	0.6%	1.5%	84.7%
GOOG	11.4%	2.4%	1.4%	0.6%	5.9%	78.2%
INTC	30.1%	1.8%	1.4%	0.2%	0.5%	66.1%
MSFT	4.9%	1.0%	1.6%	0.1%	11.7%	80.7%
NOK	12.0%	2.8%	2.3%	0.3%	6.2%	76.4%
PFE	0.7%	23.4%	0.6%	0.1%	0.4%	74.9%
WMT	7.1%	0.4%	0.6%	0.1%	5.8%	86.0%
XOM	2.5%	1.2%	0.8%	0.3%	7.3%	87.8%
YHOO	4.8%	3.4%	1.4%	0.4%	1.4%	88.6%
Average	7.9%±8.5%	3.7%±6.6%	1.1%±0.7%	0.7%±1.6%	4.0%±3.7%	82.7%±8.4%

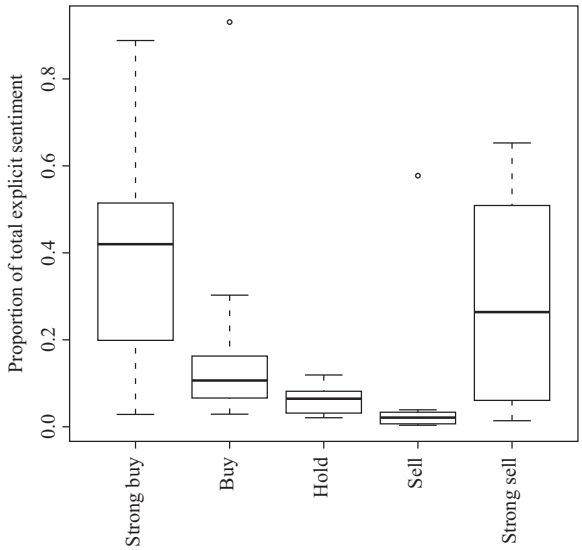


Fig. 9. Boxplot showing the explicit sentiment (when given) as a percentage of total sentiment for the 11 stocks.

To accomplish this, we first needed to evaluate and label a samplingof our posts. We used Amazon Mechanical Turk, a service to "crowdsource" labor intensive tasks. The service is a marketplace where requesters can post tasks and workers, known as Turkers, can complete the tasks for small sums of money.

Three Turkers, all from the United States, read 1976 randomly sampled posts, or roughly 3% of the total, and were instructed to label them as either "bullish" (positive outlook), "neutral", "bearish" (negative outlook), or "off-topic." A simple majority across the three evaluators was used, otherwise the post was discarded. As explained in (Ipeirotis, Provost, & Wang, 2010) the quality of Turkers varies and to ensure high quality, a preliminary labeling of 100 sampled posts by one of the authors found roughly 70% to be off-topic. Turkers who labeled posts inconsistent with this were re-examined to determine if their work was credible. In addition, we found that requiring Turkers to explain their labeling of a post with a one sentence explanation improved the quality of work.

A total of 325 Turkers were paid to label the 1976 posts with each on average completing $20[\pm 56]$ tasks. The workers were paid $\$0.03$ for the completion of both tasks (label and explain). There was perfect

Table 4
Comparing the poster's given sentiment with the Turker's sentiment (agreement is in bold)

		Turker's true message sentiment				
		Bullish	Neutral	Bearish	Off-topic	
Poster's attached explicit sentiment	Strong buy	**62**	26	7	38	133
	Buy	**26**	7	1	13	47
	Hold	3	**11**	0	12	26
	Sell	2	1	**42**	49	94
	Strong sell	2	4	**23**	36	65
	Not included	95	124	87	1006	1312
		190	173	160	1154	1677

agreement among the three evaluators on 39.12% of the posts and 84.87% of the posts received at least majority agreement. Posts with no agreement were discarded from the training set; this left 1677 posts with at least majority agreement among the Turkers.

In Table 4, the evaluator's determined "true" sentiment of the posts is compared with the poster's own attached explicit sentiment, if provided. Posts where the evaluator's sentiment label agreed with the underlying poster's explicit sentiment are highlighted in bold. This table shows a 44.9% chance that a post is identified as consistent with the underlying poster's explicit sentiment (the Turkers and poster agree). Additionally the probability that the message is evaluated as bullish when the underlying sentiment is a "strong buy" or "buy" is 48.9%. The probability of a post being evaluated as neutral when the poster provides an explicit sentiment of "hold" is 42.3%. Lastly the probability of a post being bearish given the poster has given a sentiment disclosure of "strong sell" or "sell" is 40.9%. This demonstrates that posters are providing explicit sentiment inconsistent with the written post. Of the messages that contain explicit sentiment, 40.5% are classified by the evaluators as off-topic.

In addition, 4.4% of the posts where the poster has attached an explicit sentiment of "strong buy" or "buy", were evaluated by the Turkers as actually being a bearish posts. Likewise, 2.5% of the posts where the poster gave an explicit sentiment of "strong sell" or "sell", were given a true sentiment of bullish by the Turker evaluators. The user provided explicit sentiment is consistent with the true sentiment of the messages in less than half of the posts examined.

2.2.5. *High levels of off-topic/spam posts*

Our dataset contains high levels of off-topic posts, with 68.8% of randomly sampled postings classified as such. A message is labeled off-topic when the post is clearly unrelated to the the message board and unrelated to the stock market discussion. Surprisingly, of all the off-topic posts the evaluators marked, very few were found to contain advertisements. Instead, the spam was largely political discussion and/or personal attacks on other users. Four examples of these off-topic posts are shown below:

- "Very consistent with Obama's stated goal of wealth redistribution. By 'wealth', he means anyone with over $10,000 in net worth. He wants that redistributed to his democrap people too."
- "Killing old people is Obamas answer to save Medicare and social security. Obama-care will deny old people the care they need in the last years of their life."
- "You can't blame Bush when Chenny did his thinking for him."
- "I told you to stop acting stooooooopid, ya moron. SHUT UP, SIT DOWN!"

The high level of off-topic posts within the Yahoo Finance message boards raises questions about the value of the boards themselves.

3. Classification of sentiment

3.1. *Examining meta-feature priors*

Only 17.3% of posts included user provided explicit sentiment, and of these posts, our evaluators agreed with the posters underlying sentiment in 44.9% of cases. To examine the post dynamics further, we use five additional meta-features that have previously (Castillo, Mendoza, & Poblete, 2011) been found successful in sentiment detection and compare our evaluators' labeling of the messages to determine if any interesting comparisons can be found. These features include the explicit sentiment (if any) associated with the post, if the post was a "reply"

Table 5
Prior probabilities of being evaluated in a specific class (a star "*" represent instances of less than 5) according to post features

Feature	Factor	Probability of post being labeled as:			
		Bullish	Neutral	Bearish	Off-topic
Total (Benchmark)		11.3%	10.3%	9.5%	68.8%
Explicit	Strong buy	46.6%	19.5%	5.3%	28.6%
sentiment	Buy	55.3%	14.9%	*2.1%	27.7%
	Hold	*11.5%	42.3%	*0.0%	46.2%
	Sell	*2.1%	*1.1%	44.7%	52.1%
	Strong sell	*3.1%	*6.2%	35.4%	55.4%
	Not included	7.2%	9.5%	6.6%	76.7%
Is this a "reply" to	No	13.1%	13.1%	11.9%	61.9%
another post?	Yes	10.3%	8.6%	8.1%	73.0%
n words in posts	$n \leq 3$	12.3%	7.1%	8.4%	72.3%
	$3 < n \leq 8$	9.9%	13.9%	6.2%	70.0%
	$8 < n \leq 20$	8.4%	8.4%	7.5%	75.6%
	$20 < n \leq 45$	16.1%	9.8%	7.6%	66.5%
	$45 < n \leq 159$	10.0%	9.4%	16.8%	63.7%
	$159 < n$	10.8%	21.6%	13.5%	54.1%
Posted during	No	8.8%	8.9%	5.5%	76.8%
market hours?	Yes	13.1%	11.3%	12.4%	63.2%
Post contain URL?	No	11.3%	10.3%	9.6%	68.8%
	Yes	11.3%	10.6%	8.5%	69.5%

to another post, the number of words, if it was posted during market hours, and if the post included a URL. Table 5 includes these features along with the prior probability of a message being labeled as either "bullish", "bearish", or "off-topic" by the evaluators.

As mentioned in Section 2.2.2, we theorize that messages posted during the evening could be novice individual traders conversing on the message boards during the evening when they are off work. Examining the 1677 posts that were manually classified by the Amazon Turkers, we find that a larger percentage of off-topic messages were found outside of market hours (76.8%), then during (63.2%). Using a chi-square test of independence we find statistically larger amounts of off-topic posts when the market is closed (p-value $= 4.994e$-09) when measured as a percentage of messages written. As a percentage of spam through the day, this can be seen in Figure 10. Additionally, from Table 5 messages posted during the trading day were almost twice as likely to contain sentiment of bullish or bearish (25.5%) than messages posted outside of trading hours (14.3%).

In Castillo et al. (2011), the authors found URL's to be an ideal feature to find credibility within Twitter posts. However, in our research, we found no discernible difference in using URLs to find on or off-topic posts. Furthermore, the probability of a post being labeled as bullish, neutral, or bearish is not statistically significant.

3.2. Supervised learning

In this section six different text classification algorithms are examined and compared to predict the unknown sentiment. In three separate experiments we compare model predictability using the text of the posts ("bag of words") as a feature set, the meta-data feature set, and then a combined "bag of words" and meta-data feature set. The objective is to find posts that would provide bullish or bearish sentiment that would influence the reader of that board. This becomes a three-class problem with classes of bullish, bearish, and a combined neutral and off-topic posts.

3.2.1. Representation of data

Message board posts can be represented as a simple vector of words or terms. The first step in building the learning algorithm is to break these posts into a set of terms, called the "bag of words" approach, which treats the words as a set of features within

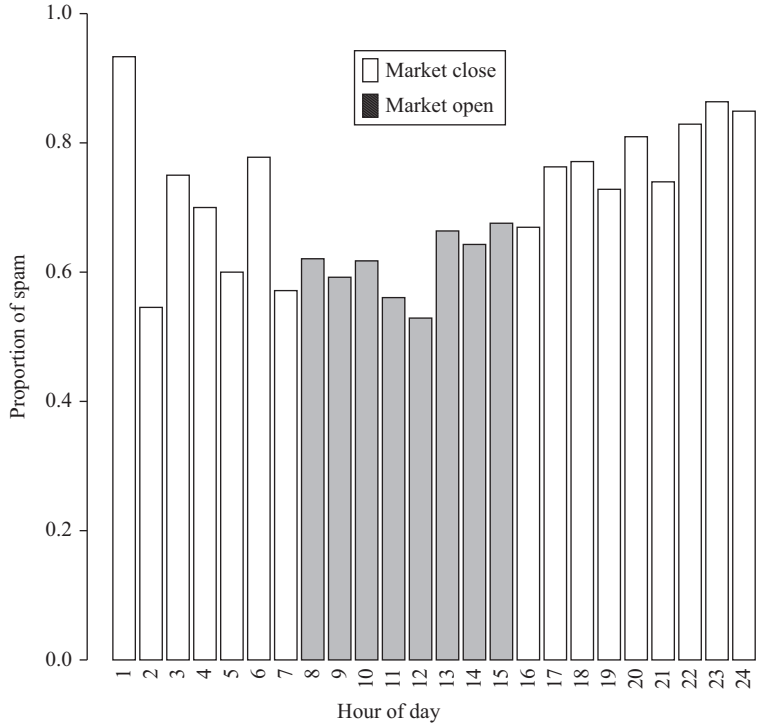


Fig. 10. Distribution of off-topic posts throughout the day.

the model. Stopwords, or frequently occurring words that add little to the meaning of the sentence (i.e. *a, by, for, in, is, or, the, to*), were removed and Porter's stemming algorithm was used to reduce the dataset further. Stemming breaks words to their stems or roots, which reduces the size of the indexing structure and may increase recall. A decrease in precision may occur however, when words with different meaning are reduced to the same stem, but in our model, better results were found with stemming enabled.

Language models often estimate a word independent of the surrounding words with the order in which the word appears irrelevant. In our model, to keep modest contextual and semantic information, we used an n-gram approach with n being 2. For example, the phrase "stocks going down fast" would be split into terms "stocks_going", "going_down" and "down_fast" with an underscore combining the words. While this increases the size of the feature set, in testing the model, better results were found. Improved performance was not found beyond 2. To offset this increased size in the dataset, and to create a more realistic model with greater generalizability (lower over-fitting), terms that appeared in two documents or less were eliminated.

Weights are assigned according to their importance to the particular document, using the popular TF-IDF scheme. The term frequency (TF) is the number of times a words appears within the posts, while the inverse document frequency (IDF) is the log of the number of posts divided by the number of posts that contain the word (DF). The TF is then multiplied by the IDF as seen in Equation 1.

$$\text{weight} = tf \times \log \frac{N}{df} \tag{1}$$

Thus the TF-IDF increases proportionately to the number of times the word appears in the posts, yet is offset by appearance of the word in the corpus. The higher the TF-IDF, the more important that word is to the post in differentiating it from other posts.

Our use of text *along with* the posts' meta-data features (Table 5) is a different approach to analyzing poster sentiment than many existing papers. For example, in Mizumoto (2012) the authors use a simple dictionary based approach to assign polarity to certain words within posts and arrive at an aggregate post polarity score. An example given in the paper is: "The design is very good and its function is superior to

the others. But the price is high." For this sentence the words "good" and "superior" have a *positive* dictionary polarity and the word "high" receives a *negative* dictionary polarity. Their method therefore gives the sentence an overall positive *polarity*. This is an interesting method, and one similar to Das and Chen (2007), but as was shown previously, some of the meta features provide high levels of separation among the sentiment classes. This is knowledge that may help with sentiment classification that is being left out of those papers.

3.2.2. Learning algorithms

Five popular supervised learning algorithms for text classification were used in this paper, the Support Vector Machine (SVM), Naive Bayes, decision tree with boosting, decision table, and the k-Nearest Neighbor (kNN). Lastly, a simple majority vote among the five algorithms is examined, i.e. three of the five algorithms should agree on the posts classification. In addition, to overcome the problem of having an unbalanced dataset, stratified sampling is employed.

The first classifier, the SVM, has long been recognized as being able to efficiently handle high-dimensional data and has been shown to perform well on text classification. The classifier is fed with pre-labeled posts and meta-data and by selecting points as support vectors, the SVM searches for a hyperplane that maximizes the margin. After training, a prediction model is built to make predictions for the new incoming data. The SVM is a two-class classifier that works with our three class problem by making multiple binary classifications (one-versus-one between every pair of classes).

The second classifier, the Naive Bayes algorithm, is an efficient probabilistic model that examines the likelihood of features appearing in the predicted classes. The *naive* aspect of the algorithm is the assumption of word independence, or that the conditional probability of a word given a class is independent from the conditional probabilities of other words given that class. The algorithm works by scanning the training data once to estimate the probabilities required for classification, and can be updated easily when new data comes in because the probabilities can be revisited with the new information.

The third classifier, a decision tree using a boosting ensemble method, has shown good results with text classification (Apté, Damerau, & Weiss, 1997). The J48 open-source version of the C4.5 decision tree was used with the popular AdaBoost variation of the boosting algorithm. For the continuous term weight attributes calculated by the TF-IDF algorithm, the decision tree algorithm considers all possible split positions and selects the one that produces the best partition. Boosting is an ensemble method that uses a weighted training set to improve model performance during a series of iterations. At the start of the training iterations, each training example begins with a weight of one. At the end of each iteration, the algorithm places more weight on misclassified examples and less weight on correctly identified ones to be used in the next iteration. Schaphire (1990) proved that it is theoretically possible to boost a weak classifier that performs slightly better than random into one that achieves arbitrary performance.

The fourth algorithm, the decision table classifier, is built on the conceptual idea of a lookup table. The classifier returns the majority class of the training set if the decision table (lookup table) cell matching the new instance is empty. In certain datasets, classification performance has been found to be higher using decision tables than more complicated models. A further description can be found in (Kohavi & Sommerfield, 1998; Knobbe & Ho, 2006; Kohavi, 1995).

The fifth algorithm, the kNN, takes the most frequent class as measured by the weighted euclidean distance among the k set of training examples in the dimension space. The k is optimized in our model using a genetic algorithm, as $k = 1$ is often not sufficient due to noise and outliers in the dataset. According to (Yang & Liu, 1999) the kNN has been shown to work as well as more complicated models. A downside to this model is the slow classification times.

The final method uses a simple majority vote among the five algorithms. It is often not clear as to which of the classifiers are optimal for the particular problem, given that are a large pool of classifiers tested. The simple choice is to choose the classifier that maximizes the chosen performance metric on the cross validation of the training set, but it does not guarantee optimal performance. Voting among multiple classifiers has been shown to sometimes increase classification accuracy (Jain, Ginwala, & Aslandogan, 2004).

3.2.3. Metrics examined

Model performance is evaluated using precision, recall, F-measure, and the kappa statistic. Precision and recall are more popular measure of performance in text classification and the results are often compared

with one another. The precision is the number of correctly identified positive examples divided by the total number of examples that are classified as positive. The recall is the number of correctly classified positive examples divided by the total number of true positive examples in the test set. Precision and recall are often achieved at the expense of the other, i.e. high precision is achieved at the expense of recall and high recall is achieved at the expense of precision. An ideal model would have both high recall and high precision. The F-measure, which can be seen in Equation 2, is provided which is a harmonic measure of precision and recall in a single measurement.

$$F = \frac{2(\text{precision})(\text{recall})}{\text{precision} + \text{recall}} \tag{2}$$

Accuracy, while it is high for our models (ranges from 67.26% to 81.75%), is not a good measure of model performance and therefore not included for our evaluation. In an unbalanced dataset, a model may misidentify all positive classes and still have high levels of accuracy; pure randomness is not taken into account with the accuracy metric. The receiver operating characteristic (ROC curve) is a common alternative to accuracy with binary classification problems, with the area under the ROC curve (AUC) used as a measure to compare different classifiers. While the AUC can be adapted for use with or on multi-class problems, the result is not as intuitive. An alternative is Cohen's kappa statistic, which takes into consideration randomness of the class and provides an intuitive result. The metric can be observed in Equation 3 where P_0 is the total agreement probability and P_c is the agreement probability which is due to chance.

$$\kappa = \frac{P_0 - P_c}{1 - P_c} \tag{3}$$

The kappa statistic is constrained to the interval $[-1, 1]$, with a kappa $\kappa = 0$ meaning that agreement is equal to random chance, and a kappa κ equaling 1 and -1 meaning perfect agreement and perfect disagreement respectively (Kaymak, Ben-David, & Potharst, 2012; Ben-David, 2008).

3.2.4. Performance

Three groups of models were built: bag of words only, meta-data only, and a combined bag of words with meta-data. A casual observation of the results found in Tables 6, 7, and 8 finds that the models easily identify the neutral/off-topic class, however considering this is 79.1% of the data, this is to be expected. This bullish and bearish classes are of

Table 6
Model performance using 10-fold cross validation where the features used were the "bag of words" only

Algorithm	Class	Precision	Recall	F-measure	Kappa
SVM	Bullish	42.65%	30.53%	35.59%	
	Bearish	36.89%	23.75%	28.90%	0.276[±0.098]
	Neutral/Off-topic	84.08%	91.11%	87.45%	
Naive Bayes	Bullish	21.61%	26.84%	23.84%	
	Bearish	19.89%	23.12%	21.38%	0.140[±0.069]
	Neutral/Off-topic	82.87%	78.37%	80.56%	
Boosted	Bullish	49.12%	29.47%	36.84%	
Decision Tree	Bearish	39.13%	16.88%	23.59%	0.296[±0.093]
	Neutral/Off-topic	84.34%	94.95%	89.33%	
Decision	Bullish	49.12%	29.47%	36.84%	
Table	Bearish	39.71%	16.88%	23.69%	0.293[±0.051]
	Neutral/Off-topic	84.21%	94.88%	89.23%	
kNN	Bullish	83.33%	2.63%	5.10%	
($k = 3$)	Bearish	100.00%	1.88%	3.69%	0.036[±0.040]
	Neutral/Off-topic	79.50%	99.92%	88.55%	
Majority Vote	Bullish	63.89%	12.11%	20.36%	
	Bearish	60.00%	1.88%	3.65%	0.118[±0.072]
	Neutral/Off-topic	80.62%	99.40%	89.03%	

Table 7
Model performance using 10-fold cross validation where the features used were the meta-features only

Algorithm	Class	Precision	Recall	F-measure	Kappa
SVM	Bullish	43.90%	18.95%	26.47%	
	Bearish	0.00%	0.00%	0.00%	0.112[±0.057]
	Neutral/Off-topic	80.44%	96.68%	87.82%	
Naive Bayes	Bullish	52.94%	28.42%	36.99%	
	Bearish	42.24%	30.63%	35.51%	0.284[±0.061]
	Neutral/Off-topic	83.41%	91.71%	87.36%	
Boosted	Bullish	46.60%	25.26%	32.76%	
Decision Tree	Bearish	45.83%	20.62%	28.44%	0.242[±0.073]
	Neutral/Off-topic	82.62%	93.52%	87.73%	
Decision	Bullish	51.22%	22.11%	30.89%	
Table	Bearish	60.00%	7.50%	13.33%	0.187[±0.096]
	Neutral/Off-topic	81.52%	96.76%	88.49%	
kNN	Bullish	49.47%	24.74%	32.98%	
($k = 22$)	Bearish	39.39%	24.38%	30.12%	0.255[±0.087]
	Neutral/Off-topic	83.07%	92.84%	87.68%	
Majority Vote	Bullish	59.46%	11.58%	19.38%	
	Bearish	50.00%	3.75%	6.98%	0.105[±0.072]
	Neutral/Off-topic	80.28%	98.49%	88.46%	

Table 8
Model performance using 10-fold cross validation where the features used were the "bag of words" and additional meta-features

Algorithm	Class	Precision	Recall	F-measure	Kappa
SVM	Bullish	50.65%	41.05%	45.35%	
	Bearish	45.00%	28.12%	34.61%	0.351[±0.073]
	Neutral/Off-topic	85.45%	91.64%	88.44%	
Naive Bayes	Bullish	21.61%	26.84%	23.94%	
	Bearish	19.89%	23.12%	21.38%	0.140[±0.069]
	Neutral/Off-topic	82.87%	78.37%	80.56%	
Boosted	Bullish	58.22%	44.74%	50.60%	
Decision Tree	Bearish	56.60%	18.75%	28.17%	0.371[±0.037]
	Neutral/Off-topic	84.98%	94.65%	89.55%	
Decision	Bullish	56.99%	27.89%	37.45%	
Table	Bearish	41.67%	18.75%	25.86%	0.301[±0.059]
	Neutral/Off-topic	84.06%	95.78%	89.54%	
kNN	Bullish	55.56%	18.42%	27.67%	
($k = 3$)	Bearish	41.38%	7.50%	12.70%	0.171[±0.047]
	Neutral/Off-topic	81.26%	97.06%	88.46%	
Majority Vote	Bullish	77.78%	11.05%	19.35%	
	Bearish	90.91%	6.25%	11.70%	0.137[±0.080]
	Neutral/Off-topic	80.72%	99.70%	89.21%	

particular interest considering this is what influences others to act in the market. The class distribution of these two classes are 11.3% and 9.5% respectively.

Examining the "bag of words" model in Table 6, it can be observed that boosted decision tree performs the best according to the kappa statistic at $0.296[\pm 0.093]$, but it is not statistically higher than the Support Vector Machine or the decision table. The kNN with a k optimized at 3 performs significantly worse at a kappa of $0.036[\pm 0.040]$. Additionally, the F-measure for the top three models range from 35.59% to 36.84% for the bullish class and from 23.59% to 28.90% for the bearish class. The best combination among the models appears to be the SVM with a F-measure of 35.59% for bullish and 28.90% for bearish. This model identifies 30.53% of all relevant bullish posts and correctly classifies 42.65% of all attempted bullish post. Likewise for bearish posts the SVM identifies 23.75% of all relevant bearish posts and correctly identifies 36.89% of all attempted bearish posts.

In Table 7 the meta-features are used as the only set of features in the model. In this model, the Naive Bayes model has the highest mean kappa statistic at $0.284[\pm 0.061]$, but this is not statistically higher than the boosted decision tree or the kNN. Since the meta-data was discretized to allow for comparisons for all six models, information was lost which would have possibly allowed the SVM to have greater performance. The SVM is able to be applied to categorical variables by creating dummy variables for each categorical attribute value presented in the data. While this allows the other models to work, such as the Naive Bayes, it loses information in the dataset with which the SVM generally excels. This is a reason why the SVM works so well with the "bag of words" featureset; the words are represented as numerical weights from TF-IDF calculations. Additionally the SVM performed poorly on classifying the bearish instances; it classified all as spam. The kNN with a genetic algorithm optimizing k at 22 showed significant improvements from using the "bag of words" featureset. The best performing model from the meta-data featureset, the Naive Bayes, performed as well as the best performing model from the "bag of words" featureset, the boosted decision tree.

The combined "bag of words" and meta-features featureset can be seen in Table 8. While the boosted decision tree has the highest kappa, it is not statistically superior to the SVM ($0.371[\pm 0.037]$ vs $0.351[\pm 0.073]$). However, the boosted decision tree is significantly better than the best performing models from the other two datasets; the SVM is statistically the same as the best performing models. We conclude therefore that the boosted decision tree is the best performing model when examining the kappa statistic. Examining the F-measure, the decision is split, with the boosted decision tree having a higher metric for bullish sentiment while the SVM has higher performance for the bearish sentiment.

In all six of the models, using all three of the featuresets, both the precision and recall are higher for the bullish post than for the bearish posts. This may be related to the Amazon Turkers having difficulty in evaluating posts that are often a fine line between a bearish posts and a rant. The low performance levels of the models demonstrates the difficulty in performing sentiment analysis on frequently poorly written, ambiguous, online posts. As noted in Das and Chen (2007) message board posts sentiment analysis is more difficult than creating spam filters, where the characteristics of spam versus non-spam emails are distinct. The difference between a bullish versus neutral/spam posts and a bearish versus neutral/off-topic posts is often subtle resulting in high levels of false positives. However, anytime humans are involved, results can vary. We tried to minimize this as much as possible by following ideas from Ipeirotis et al. (2010) by paying three Mechanical Turkers to analyze posts and then using majority voting to arrive at the evaluation. We additionally found that requiring Turkers to explain their labeling of a post with a one sentence explanation and also by confining the workers to the United States improved the quality of work. Also, all of the messages included in the training set were read over by one of the authors (who worked seven years as a trader at the Chicago Stock Exchange) to provide an extra layer of assurance.

3.2.5. Examining important features

In Table 9 the χ^2 statistic for the top features are examined for the three models. This statistic evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class.

In the bag of words model, the top two words are the appropriate "buy" and "sell." The n-gram "bristol_myers" is the name of one of the stocks covered in our dataset and it is observed to be used 50% bullish. Also to no surprise, the term "Obama" has a χ^2 of 48.9 and is observed to be 78% in the neutral/spam class – post mentioning the President is political in nature and not often related to a

Table 9
Assessing χ^2 for the top 15 features of the three models

"Bag of words" only	χ^2	Meta-features model (includes 11 features only)	χ^2	"Bag of words" and meta-features model	χ^2
buy	172.4	Explicit sentiment	486.3	Explicit sentiment (meta)	486.3
sell	115.1	Count of the number of days	87.7	buy (word)	172.4
bristol_myers	74.5	the author has been active		sell (word)	115.1
csco	68.4	Is this a suspicious user?	44.2	Number of days the author	87.7
stock	63.6	via Levenshtein metric		has been active (meta)	
market	62.6	n words in post	43.7	britol_myers	74.5
asia	51.4	Avg. number of posts by user	43.4	csco	68.4
Obama	48.9	per active day		stock	63.7
strategy	42.8	Length poster has been	40.1	market	62.6
level	40.9	a Yahoo member		asia	51.4
nokia	40.8	Posted during market hours?	33.5	Obama	48.9
metal	39.9	Maximum number of post	32.9	Is this a suspicious user?	44.2
cheaper	38.3	by author in an active day		Levenshtein metric (meta)	
investor	38.1	Is this a suspicious user?	30.3	n words in post (meta)	43.7
oil_asia	38.0	via Cosine similarity		strategy	42.8
		Is this a reply to another post?	11.1	Avg. number of posts by	42.4
		Post contain URL?	0.2	user per active day (meta)	

particular stock. Examining the top meta-features in the other two models, the "explicit sentiment" feature has the largest χ^2 at 486.3, followed by the "Count the number of days the author has been active", and third the Levenshtein edit distance metric feature. As explained earlier in this paper, in 44.9% of the instances in which explicit sentiment is given, the true sentiment is identical to the explicit. This provides a high level of separation for the classes and is an ideal feature to include in the model.

4. Prediction of stock price

4.1. Introduction

In this section we use the stock's predicted sentiment and the user-provided explicit sentiment along with the post count, volatility, and stock market movement over n days for training in a neural network. The features are optimized with a genetic algorithm and the model is then used to predict the stock's future direction of "up" or "down." In theory, message board posts can influence board participants, and since these individuals can affect the stock market, we intend to discover if this relationship is measurable against a benchmark that included historical price data only. The methodology and results follow.

4.2. Data

The best performing sentiment model, the boosted decision tree, is used to find the sentiment from the remaining $66,172$ posts. This predicted sentiment, along with the user-provided explicit sentiment and post counts are aggregated for each stock from the close of the previous trading day $t-1$ until the close of trading day t. For each of the attributes, three calculations create features which are then lagged to capture the underlying trend. The first calculation is the log value of the attribute for $t-n$ where n is $0, 1, 2, 3$. The second feature is the percentage change of the attribute from time t for $t-n$ where n is $1, 2, 3$. Lastly a change in the direction of the value at time t is compared to the value at time $t-n$ with 1 equaling "up" and 0 equaling "down." The feature calculations are shown in Table 10. The attributes' descriptions appear next.

Price attribute Price is the daily closing price of the stock adjusted for dividends.

Volatility attribute The stock's intraday volatility is the variation of the price over time and is calculated

Table 10
The log attribute, percentage change from $t - n$, and the binary feature indicating an "up" or "down" movement in attribute direction from t to $t - n$

Type	Description
Continuous	$\log(\text{attribute})_{t-n}$ where $n = 0, 1, 2, 3$
Continuous	$(\text{attribute}_t - \text{attribute}_{t-n})/\text{attribute}_{t-n}$ where $n = 1, 2, 3$
Binary	$\text{attribute}_t > \text{attribute}_{t-n}$ where $n = 1, 2, 3$

daily for each stock by subtracting the stock's daily low price from the stock's daily high price and dividing over the average of the daily opening and closing prices. There are multiple ways of calculating this, but we used the method explained in (Das & Chen, 2007).

Post count attribute This is the total number of posts that are provided for a particular stock for time t.

Predicted sentiment attribute (bullish and bearish) This is the predicted sentiment which was obtained by using our best classifier, the boosted J48 decision tree, by using the "bag of words" and meta-features. The third class that was predicted (neutral/spam) is not given since the sentiment should not influence traders.

Explicit sentiment attribute (buy and sell) This is the user provided sentiment which can be selected from Figure 1. "Strong buy" and "buy" were combined into "buy", and "strong sell" and "sell" were combined into "sell."

4.3. Classification

An artificial neural network (ANN) was used for the classifier to predict the future up or down movements in the stock price. Studies provide some evidence that nonlinear models are able to produce better predictive results and of these models, the ANN has provided strong results for classifying financial instruments. Additionally, ANN works well where the data contains high amounts of noise or information is missing (Yu, Wang, & Lai, 2005; Yoon & Swales, 1991).

Specifically the ANN was a feed-forward ANN trained by a back-propagation algorithm with one hidden layer of size n calculated from the formula in Equation 4. Favorable findings were found using 500 training cycles with a learning rate of 0.3 and a momentum of 0.2.

$$n = \frac{(\text{number of features} + \text{number of classes})}{2} + 1 \tag{4}$$

The posts' data that was obtained for the eleven stocks was retrieved during different timespans (see Table 1). In addition, the observed stocks have high correlations among one another with correlations for price as high as 0.947 and post count correlations as high as 0.710. Using cross-validation with the ANN in predicting the individual stocks up or down movement and obtaining a variance of performance, would therefore not be appropriate. Instead, bootstrap sampling was done with the training set comprised 75% of the data and the test set made up of the remaining 25%. Twenty new training sets were bootstrapped with replacement and tested using the ANN on the test set.

4.4. Feature selection and performance measures

A genetic algorithm (GA) is used for feature selection of classifier input variables in order to reduce the model complexity and increase the model interpretability. Feature selection is a process of selecting a subset of the original data to reduce feature redundancy or to eliminate features with little information while maintaining an acceptable classification performance. An additional benefit to finding the "optimal" subset of features from a larger subset of features is the reduction of training time of the classifier. The GA is especially suitable for this task because it performs a randomized search and is not susceptible to getting stuck in a local minima (Kim & Street, 2004; Vafaie & De Jong, 1992; Jarmulak & Craw, 1999). Principal component analysis (PCA), a popular dimensionality reduction algorithm, is not appropriate in this case because PCA does not consider the relationship between the response and other input variables in the process of data reduction.

The GA population begins with five randomly generated vectors of binary genes (the chromosomes), with relevant features represented as "1." Selection is accomplished via tournament, with two chromosomes chosen at random and entered into a *tournament* against each other. The classifier trains and tests on both chromosomes separately and returns an evaluation function, which is accuracy in our balanced testing set. The best performing, according to accuracy, is chosen to be used as the parent. A tournament is done a second time to get another parent. These parents are then combined via a uniform crossover operator of probability of 0.50 to create a new, hopefully better, offspring. Mutation of 3% was used to add diversity into the population and to ensure that it is possible

to explore a greater percentage of the search space. Lastly, thirty generations were used before the process completed.

In measuring performance, accuracy is included since the class distribution is nearly even. In the test set, price_t is greater than price_{t+1} in 44 out of 84 observations (the "up" class), or 52.38% of observations. In this experiment, we are testing to determine if either the explicit or predicted sentiment is more predictive of the market. Both are compared against the baseline performance measure, which is defined as the model built using features calculated from the price and volatility attributes.

The F-measure, which was included in the prediction of sentiment, is not included since the precision and recall are not equally important in the prediction of "up" or "down" movements. Additionally ROC is not included because of its inability to handle example-specific costs. For example, precision in a trading model is more important than missing a trading opportunity. This can be compared to a bank accepting or denying a credit card transaction based on the probability of it being a fraudulent transaction. Approving a fraudulent transaction would amount to the credit card company losing the transaction cost, while denying a legitimate transaction amounts to a relatively small amount and a slightly annoyed customer (Fawcett, 2004). This is similar to trading in that missing an opportunity to trade (recall) is small compared to the loss of capital for incorrectly identifying the trade (precision).

4.5. Results

Performance of the three models compared to the baseline for $t + 1$ days can be seen in Table 11. The baseline is an optimized model containing only the price and volatility attribute features and has an accuracy of $56.67\%[\pm 4.33\%]$. The model containing all the features (price and volatility along with both predicted and explicit sentiment features) has an accuracy of $57.74\%[\pm 4.03\%]$, the model containing price and volatility along with explicit sentiment features (excludes the predicted sentiment) has an accuracy of $54.88\%[\pm 4.83\%]$, and lastly the model containing the price and volatility along with predicted sentiment features (excludes the explicit sentiment) has an accuracy of $62.86\%[\pm 2.43\%]$. The model containing the price and volatility features along with the predicted sentiment performs statistically higher than the model containing the price, volatility, and explicit sentiment features. This demonstrates that the predicted sentiment is more predictive of the market direction than just including the explicit sentiment for $t + 1$ days; significance was not found at $t + 2$ or $t + 3$ days. Additionally excluding the price and volatility features resulted in no significance over the baseline.

The features included in the best performing model can be seen in Table 12. Cumulative post counts, price, price volatility, and both predicted bullish and bearish sentiment are observed to contribute to the predictability of the future price of the market. Explicit sentiment however, is not found to contribute to predictability.

5. Examining extreme posters

5.1. Who are these users?

Within the message boards, a suspicious user would be one who provides a large number of posts or posts across multiple accounts. A "pump and dumper" is the term for an individual who posts multiple messages, often through several accounts, as a means of influencing others to buy and sell stock. His or

Table 11
Performance of ANN in predicting $t + 1$

Features included		Precision	Recall	Accuracy	Stat. Sig.
price/volatility features only (baseline)	Up	59.84%	52.50%	$56.67\%[\pm 4.33\%]$	–
	Down	53.96%	61.25%		
price/volatility features, predicted sentiment, and explicit sentiment	Up	60.24%	56.82%	$57.74\%[\pm 4.03\%]$	no
	Down	55.29%	58.75%		
price/volatility features and explicit sentiment only	Up	57.49%	53.18%	$54.88\%[\pm 4.83\%]$	no
	Down	52.42%	56.75%		
price/volatility features and predicted sentiment only	Up	64.88%	64.09%	$62.86\%[\pm 2.43\%]$	yes
	Down	60.89%	61.50%		

Table 12
Features selected by the genetic algorithm for the best performing model (all features, excluding explicit sentiment)

Feature
$\log(\text{cumulative posts})_{t-n}$ where $n = 0, 1, 2, 3$
$\log(\text{price})_{t-n}$ where $n = 0, 1, 2$
$(\text{price}_t - \text{price}_{t-n})/\text{price}_{t-n}$ where $n = 2$
$\text{price}_t > \text{price}_{t-n}$ where $n = 1, 2$
$\log(\text{volatility})_{t-n}$ where $n = 1$
$(\text{price volatility}_t - \text{price volatility}_{t-n})/\text{price volatility}_{t-n}$ where $n = 1, 2$
$\log(\text{predicted bullish sentiment})_{t-n}$ where $n = 2, 3$
$(\text{predicted bullish sentiment})_t > (\text{predicted bullish sentiment})_{t-n}$ where $n = 1$
$\log(\text{predicted bullish sentiment})_{t-n}$ where $n = 0, 1, 2, 3$
$(\text{predicted bearish sentiment})_t > (\text{predicted bearish sentiment})_{t-n}$ where $n = 3$

her goal is to disseminate false information through venues such as message boards as a means of inflating or deflating the price of the stock. A second group of users, who appear to be significant on the online message boards, are those who post off-topic communications online. Often these posters create contagious chaos by attacking others; these users are of no value to individuals searching for stock related information. An example of an off-topic communication by user Chavo Ortega on the message boards posted the following off-topic message on April 18, 2012:

> You are a pathetic punk. Get off the library computer and go downtown to beg for change. I am assuming you will be dumpster diving for lunch and dinner, before using your begging money for a cheap bottle of booze. You are such a bum.

In this section we examine posters who post large amounts of messages, and those who we suspect of having multiple accounts. These suspected multiple account owners are found using two methods. The first metric that we used is the Levenshtein edit distance metric to determine username similarity among posters (Iofciu, Fankhauser, Abel, & Bischoff, 2011). In the second metric we use cosine similarity to determine if the posted messages of users have strong similarity to existing users' posts.

Lastly we examine the length of time a poster has had an account opened. The ease of opening a new account on Yahoo may lead to abuse, and with an incentive to be an influential agent on stock message boards (DeMarzo et al., 2003), we wanted to explore the relationship (if any) between the length of time an account was open and the quality of posts. Results of the following analysis can be found in Table 13.

5.2. Frequent posters

Seeing that the top posters to the individual stock boards write anywhere from 7.01% to 21.49% of all the messages, we ask if this is representative of an individual with legitimate concerns and ideas, or an individual looking to mislead others? Counting only the days on which the top poster of each stock posts messages, we find an average of $18[\pm 9]$ messages per day. While day-traders, often trading from the loneliness of their homes and perhaps seeking companionship with others online, this number still appears outside of the range of normal posting behavior.

Table 14 illustrates the average number of posts by posters on days in which they are active. Over 99% of all posters are writing less than an average of ten posts per day, with less than 1% writing eleven or more per day. Less than 0.09%, or 6 users from our dataset, are writing 26 or more posts per day. As a percentage of participation on the message board, users writing 11 or more posts contribute 19.4% of all messages while being less than 1% of active users. In our study, these users are considered outliers, or "suspicious", while those who post less are considered to be within the normal range of posting behavior.

By examining the posting behavior of these two groups, we find that the "suspicious" group provides slightly more explicit sentiment of "strong buy", "buy", "hold", "sell" and "strong sell" with their posts. This number is statistically significant[3] with 17.4% of suspicious posts and 15.6% of non-suspicious posts containing explicit sentiment.

5.3. Username similarity

Some users appeared to be posting under multiple usernames. Examples found included, "madmilken67", "madmilker69", "madmilken69", "madrnilker69", "madmilker79", "madmilken69", "madlecher69" (*leche* is Spanish for *milk*), and "rnadrnilker". Another example is "bobbyjoe51", "bubblyjoe51", "bubbyjoe51", and "bublyjoe51." The Levenshtein edit distance metric measures the minimum number

[3] All measures marked statistically significant in this paper, unless otherwise notified, are at the 0.05 level.

Table 13
Prior probabilities of being evaluated in a specific class (a star "*" represent instances of less than 5) according to poster features

Feature	Factor	Probability of post being classified as:			
		Bullish	Neutral	Bearish	Off-topic
Total (Benchmark)		11.3%	10.3%	9.5%	68.8%
Maximum number of	up to 3	18.3%	14.0%	8.7%	59.0%
posts by author	3 to 10	11.9%	10.3%	12.3%	65.4%
in any active day	11 to 25	8.7%	6.9%	5.5%	78.8%
	26+	9.8%	12.5%	11.3%	66.4%
The number of	up to 5	17%	12%	9%	63%
days the author	6 to 10	8%	12%	7%	74%
has been active	11 to 25	15%	9%	5%	71%
	26 to 50	10%	6%	4%	80%
	51+	7%	11%	19%	63%
Average number of posts	up to 3	16.3%	13.8%	6.2%	63.7%
by user per active day	3 to 10	8.7%	6.8%	12.1%	72.5%
	11 to 25	11.3%	15.2%	11.3%	62.3%
	26+	* 1.3%	* 2.5%	* 3.8%	92.4%
Length poster has	<3 months	9.0%	13.4%	7.5%	70.1%
been a Yahoo	3 to 6 months	15.5%	12.1%	* 5.2%	67.2%
member	6 to 12 months	8.6%	* 2.9%	7.5%	81.0%
	1 year to 5 years	5.9%	11.8%	7.5%	74.7%
	5+ years	20.5%	14.2%	6.3%	59.0%
	Info. hidden	10.6%	9.8%	11.9%	67.7%
Poster has username similarity?	No	11.6%	11.0%	7.2%	70.2%
(Levenshtein metric)	Yes	10.4%	7.3%	19.2%	63.1%
Postings are similar with others?	No	15.9%	15.2%	9.3%	59.6%
(Cosine similarity above 0.70)	Yes	7.4%	6.2%	9.7%	76.7%

of edit operations required to transform $username_1$ to $username_2$. A conservative estimate to deem a username suspicious was determined to be where another username shared 75% of the total characters. This found 498 usernames out of the 6906 during the timespan that were potentially suspicious, or roughly 7% of the total[4].

While it is possible that Yahoo's *alternative username suggestion algorithm* is partially to blame for username similarities (see Figure 11), a statistical analysis comparing the mean number of posts by the usernames which had a low Levenshtein edit distance metric (the suspicious users) and the users with a high Levenshtein edit distance metric (the non-suspicious users), found statistical difference between the two groups, with 25.7 ± 99.7 and 8.6 ± 39.3 posts respectively. From Table 15, users with a high Levenshtein edit distance metric comprised 92.79% of posters and wrote 81.30% posts, while the suspicious users comprised 7.21% of posters and wrote 18.85% of posts.

Table 14
Examining users' post per active day and their contribution to the message boards

	up to 3	4 to 10	11 to 25	26+
% of total posters	87.42%	11.76%	0.74%	0.09%
% of total post written	33.37%	47.24%	17.49%	1.90%

5.4. Post similarity

Users with similar usernames were found to write a statistically greater number of messages. We now

[4]Our metric for determining if a user was suspicious was if it shared similarity with at least one other user. The majority of all usernames did not share similarity with another, however, for those who did, 70% shared similarity with 1 other, 20% with 2, and 10% with 3 or more. Our lack of data at these smaller scales did not produce results that would have generalized well, so we excluded it from this paper.

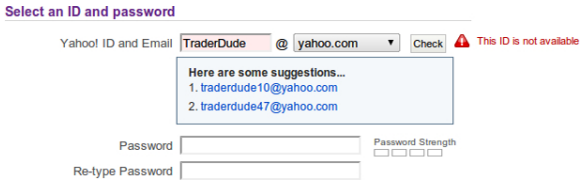


Fig. 11. Yahoo suggesting alternative usernames.

Table 15
Examining username similarity and their contribution to the message boards

	High Levenshtein (Non-suspicious poster)	Low Levenshtein (Suspicious poster)
% of total posters	92.79%	7.21%
% of total post written	81.30%	18.85%

consider the text similarity of post by different usernames. Strong similarity of post text could signify another method of finding one user potentially writing under multiple account.

A common similarity function for text is the cosine similarity, which is the measure of similarity of two vectors by measuring the cosine angle between them. The posts are represented as simple vectors of words or terms. Stopwords were removed, which are frequently occurring words that add little to the meaning of the sentence (i.e. *a, by, for, in, is, or, the, to*). The vector $\vec{V}(post_1)$, derived from the post, contains one vector component for each term in the dictionary. The vector components are calculated using TF-IDF weighting which was explained in earlier in this paper. To quantify the similarity between two posts, the cosine similarity of vector $\vec{V}(post_1)$ and $\vec{V}(post_2)$ are calculated using the formula in Equation 5.

$$\text{sim}(post_1, post_2) = \frac{\vec{V}(post_1) \cdot \vec{V}(post_2)}{|\ \vec{V}(post_1)\ ||\ \vec{V}(post_2)\ |} \quad (5)$$

The numerator represents the dot product of the vectors $\vec{V}(post_1)$ and $\vec{V}(post_2)$, and the denominator is the product of their Euclidean lengths (Manning, Raghavan, & Schütze, 2008).

There are a total of 6906 posters writing a combined $67,849$ posts. The posts for each author were combined and those whose total number of words (with stopwords omitted) amounted to less than 20 were eliminated. This prevented strong cosine similarity of users who wrote one small posts only, for example, "This stock is a strong buy." This reduced our number of examined users to 4250 and reduced our total number of examined similarities ($\frac{N(N-1)}{2}$) from 23.8 million comparisons to 9 million.

Figure 12 shows the proportion of the total number of comparisons where the cosine similarity between two posts are between sim_1 and sim_2. Over 97.5% of the post comparisons are below a cosine similarity metric of 0.40.

We mark users that have cosine similarities of above 0.70 to be deemed suspicious. This amounts to a total of 464 poster accounts with a total of 3776 connections among the group with cosine similarities above 0.70. Examining these 464 users, or 6.7% of the total, it is found that they write 52.7% of the total posts and provide 45.0% of all the user provided explicit sentiment (see Table 16). The suspicious group is also found to provide a statistically greater number of explicit sentiment that is "strong buy" (mean of 0.5 ± 4.1 per user versus 5.3 ± 29.5) and that is "strong sell" (mean of 0.3 ± 4.9 per user versus 2.29 ± 19.4).

Two examples of accounts that have high similarities with other posters follow. The poster *ball_keller, iaxaytvwqzt, kayla.frye9, kjhsdjhi, pdbdyxpafjy, rodgers.har, xludptmdetu* all appear to be the same individual with perfect cosine similarity of 1.0 among the group. The messages posted are rambling, "you should be happy the markets going up if the market just went down you wouldn't be able to trade it anymore." These seven usernames are all quite different, and the Levenshstein edit distance metric previously discussed, would not have found any similarities. An additional example of high cosine similarity among posters is the user *fred.jackson* who has very high levels of similarity with 56 other accounts. These accounts post mostly insults toward other users. The objective of this user is unknown and emails to the user have not been returned.

5.5. Length account open

Yahoo allows users to have an online profile that displays information about the user, such as the length his or her account had been opened, age, gender, and location of residence. We captured the total length of time the 6906 posters accounts had been open at the end of the study to determine if a higher than normal rate of messages came from users whose accounts had been open for shorter periods of time. Opening an account on Yahoo is easy and anonymous, so we wanted to determine if there was a correlation between the number of posts, levels of explicit sentiment,

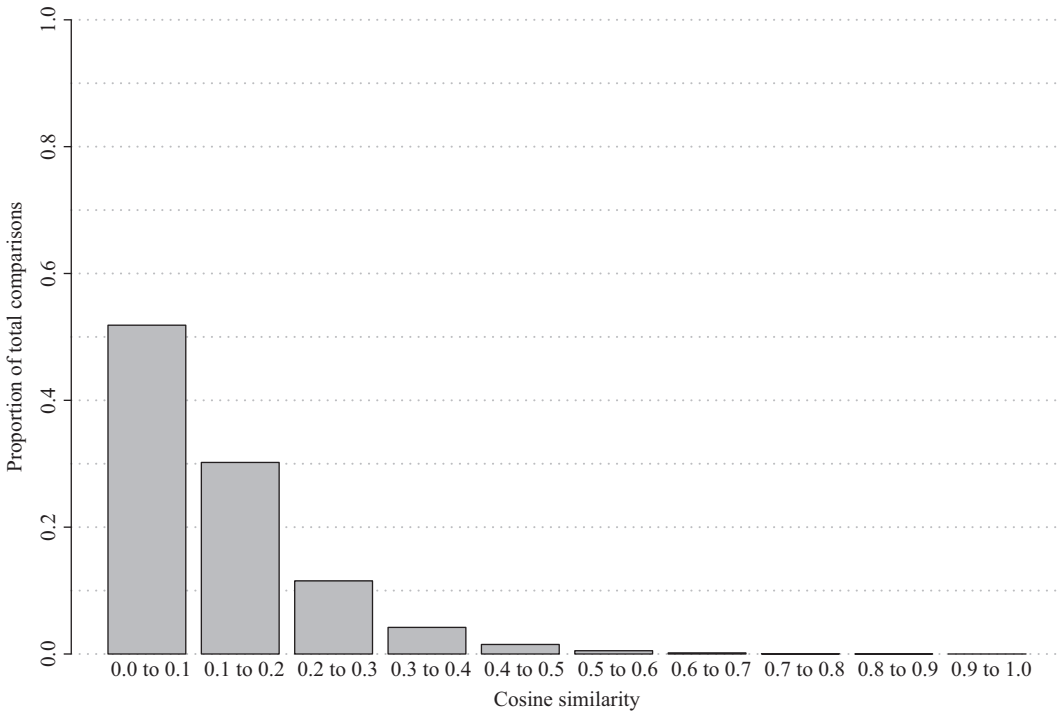


Fig. 12. Proportion of combinations where the cosine similarities fit the observed.

Table 16
Examining users that have have strong post similarity and their contribution to the message boards

	Low cosine similarity (Non-suspicious poster)	High cosine similarity (Suspicious poster)
% of total posters	93.28%	6.72%
% of total post written	47.33%	52.67%

and the length of time the account had been open. We binned users into five groups according to the length their account had been opened: less than 3 months, 3 to 6 months, 6 months to 1 year, 1 year to 5 years, and greater than 5 years. A sixth group was added for members for which we were unable to retrieve account information. A setting within the profile allows users to hide the length their accounts have been open.

By default, Yahoo account profiles are publicly available, however within our dataset, as seen in Table 17, 46.5% of the account profiles were hidden, meaning the user of the account made an explicit decision to hide the account details for whatever reason (this information is also visualized in Figure 13). While this percentage appears high, we are unable to compare this to a group of Yahoo members outside of the stock message boards. Yahoo has recently changed the access to this information and it is no longer able to be retrieved on a large scale basis. However, the authors of this paper believe anecdotally that this percentage appears high.

Examining only those posters who supply their account information, posters whose accounts had been open for less than 3 months (new users) describe 17.5% of users, yet write only 9.5% of total posts and provide only 5.9% of the total explicit sentiment. Because accounts are easy to open and the incentives high, we expected scammers to more frequently open new accounts as a means of expressing their sentiment via messages and explicit sentiment disclosure. However, this is the opposite of what was found – new users did not post as frequently. On second observation, this does appear to be consistent with literature on the behavior of individuals new to any group; new users tend to observe (e.g. lurk) until they become aware and confident of the dynamics of the group. According to Nonnecke and Preece (2000), these new users can make up over 90% of all online groups. Users in our dataset appear to become more active on the boards after six months, accounting for

Table 17
Comparing the length a poster has had their account open

Length poster has been a Yahoo member	As % of total active users	As % of total posts written	As % of total explicit sentiment provided
<3 months	17.5%	9.5%	5.9%
3 to 6 months	3.5%	3.1%	1.4%
6 months to 1 year	5.8%	6.6%	8.5%
1 year to 5 years	10.7%	10.8%	9.3%
5+ years	16.1%	17.9%	19.5%
Information hidden	46.5%	52.1%	55.4%
	100.0%	100.0%	100.0%

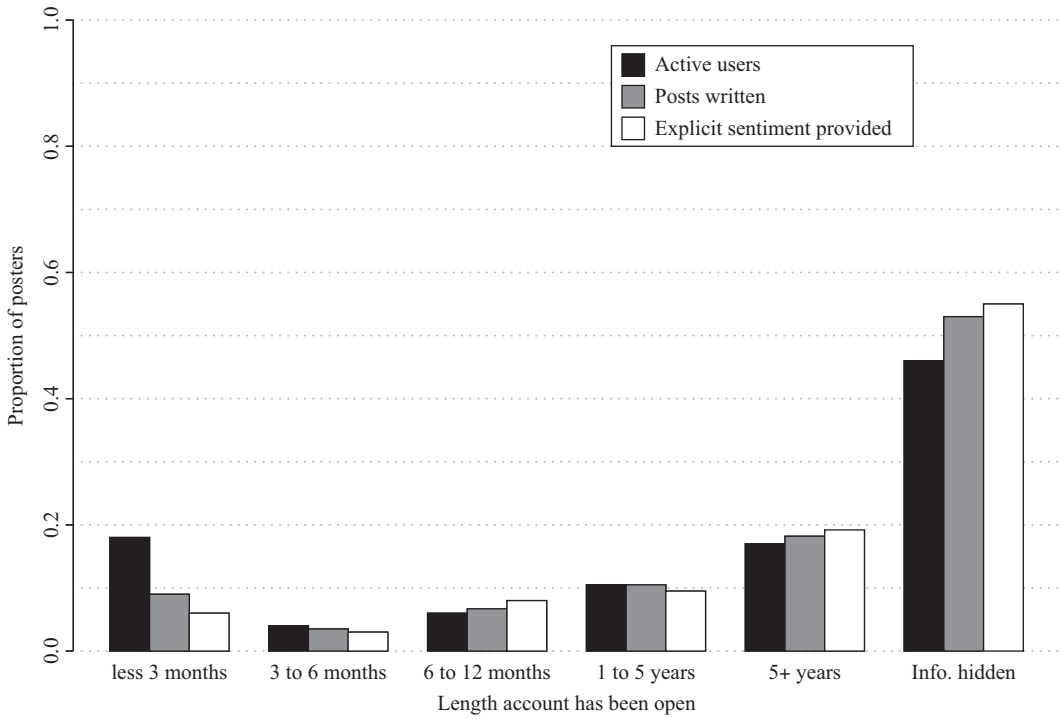


Fig. 13. Length Yahoo account had been opened as compared against total % of message written for that group.

a larger percentage of both posts written and explicit sentiment provided. Examining users based on the length of account opened to find suspicious users does not appear to be worthwhile in the message boards.

5.6. *Predictability when eliminating the "suspicious" users*

From Section 5, suspicious users are those who posts an abnormally large number of posts, or who are suspected of posting across multiple accounts. While it is not certain that these users are seeking to influence others in an unethical way, such as by touting stocks (i.e. "pump and dumpers"), their activity on the message boards raises suspicion. Citing the study by Frieder and Zittrain in (2007), where "pumpers" sent large quantities of messages in an attempt to entice others to buy, found that investors who bought lost on average 5.25%. This suggests that these users are providing sentiment opposite of their true trading decision. Our hope is that by eliminating these participants in the message boards, predictability will increase.

The first step is eliminating all users who write on average eleven or more posts per day, had a low Levenshtein edit distance metric (high similarity to existing usernames), and a high Cosine similarity of

Table 18
Performance of ANN in predicting $t+1$ with suspicious posters eliminated

Features included		Precision	Recall	Accuracy	Stat. Sig.
price/volatility features only (baseline)	Up	59.84%	52.50%	56.67%[±4.33%]	–
	Down	53.96%	61.25%		
price/volatility features, predicted sentiment, and explicit sentiment	Up	59.80%	54.09%	56.90%[±3.14%]	no
	Down	54.30%	60.00%		
price/volatility features and explicit sentiment only	Up	60.85%	52.27%	57.38%[±5.10%]	no
	Down	54.55%	63.00%		
price/volatility features and predicted sentiment only	Up	58.40%	51.36%	55.36%[±4.37%]	no
	Down	52.76%	59.75%		

posts when comparing to other users' posts. This reduces the posts count to $27,370$. Using the same feature reduction methodology as stated previously, the results for predicting the price direction can be found in Table 18. The results in Table 18 show no discernible difference over the baseline (price and price volatility only) featureset. The predicted sentiment of the suspicious users therefore was found to be predictable of the market direction for the following trading day.

The authors believe there are several possible reasons why greater predictability was not found with these users removed. First, the suspicious users may be influential enough to get others to buy and sell based on their recommendations. If the influence is strong enough, this would *move* the underlying stock price. A second possibility is that posts were reduced over 50% when excluding the suspicious users. The suspicious users often provided sentiment on days when other posters were not. For example, the average daily predicted bullish sentiment is $19.66[\pm 15.05]$ for all users, and $11.03[\pm 9.60]$ when excluding suspicious users. Average daily predicted bearish sentiment is $7.62[\pm 5.82]$ for all users, and $3.63[\pm 3.41]$ when excluding suspicious users. This lack of sentiment and diversity of sentiment appears to adversely affect predictability. Third, there is no "ground truth" as to what a suspicious user is. We have no way of knowing if these users are actually "pump and dumpers" or simply energetic posters with strong opinions.

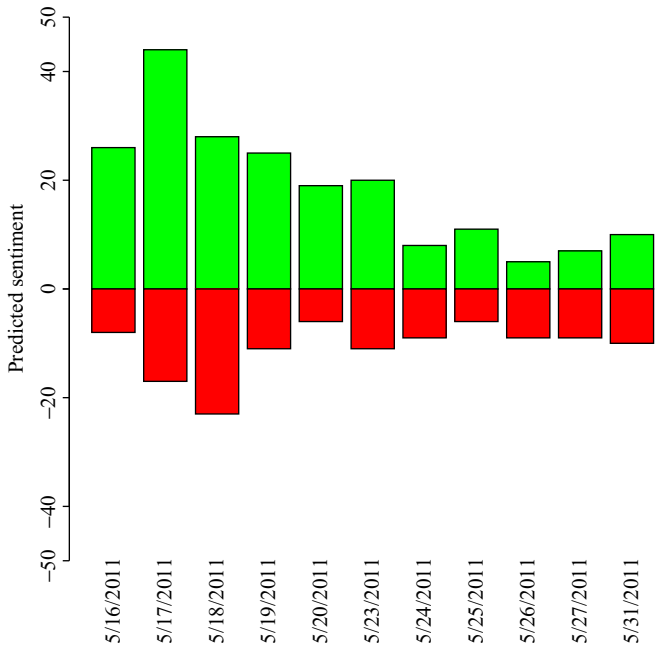
To demonstrate the reduction in sentiment when excluding suspicious users, Figure 14 examines Wal-Mart (symbol: WMT) predicted sentiment over eleven days. The graph aggregates predicted bullish sentiment in gray and bearish sentiment black. As can be seen in Figures 14a and 14b, the reduction in sentiment during this period is 71%.

6. Conclusion

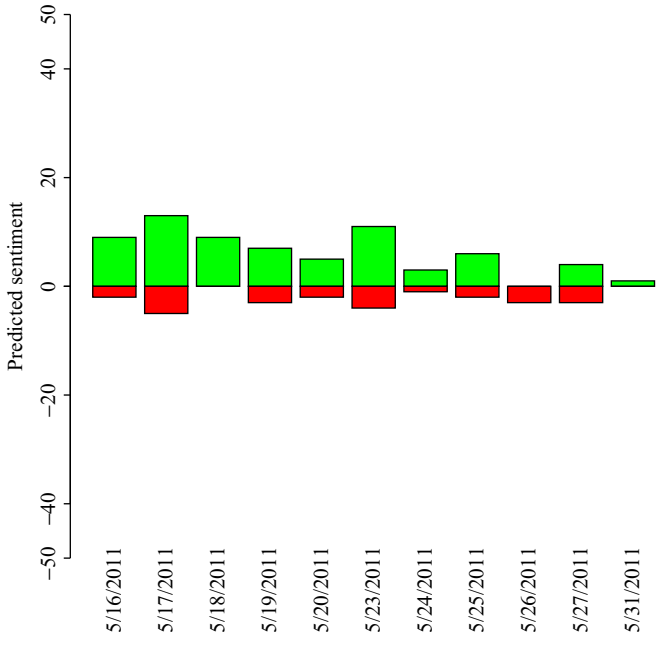
In conclusion, this paper examined the posts and participants of a popular online message board. A small 17.3% of posts included explicit sentiment of "strong buy", "buy", "hold", "sell" and "strong sell." Using a supervised text classification model, we were able to find sentiment contained in approximately double the posts. Spam was found in 68.8% of posts which raised questions about the usefulness of the boards. Additionally, certain features related to the posts and posters displayed stronger probabilities of being classified by evaluators as either "bullish", "neutral", "bearish", or "off-topic." Messages posted during market hours had lower prior probabilities of off-topic messages and had higher probabilities of sentiment than messages posted outside of market hours.

We explained that with the potential for profit in stocks and the popularity of online sentiment trading websites, certain mischievous individuals may be attempting to influence others within the message boards. These "pump and dumpers" may be undermining the online sentiment trading websites to have sentiment inline with their trading goals ("bullish" sentiment if they own the stock or "bearish" sentiment if they are short the stock and want to push the stock down). Four methodologies were examined for finding these outlier users, with two of these methods finding the existence of users that had the potential of having multiple accounts. The Levenshstein edit distance found similarity of usernames and Cosine similarity found posters with similar writings. This found posters with varying levels of sentiment and off-topic posts.

Lastly we used artificial neural networks to determine that the markets were predictable when using



(a) Predicted sentiment from all posters



(b) Predicted sentiment for non-suspicious posters only

Fig. 14. Comparison of the predicted sentiment from all posters versus the sentiment from the non-suspicious posters only for the stock Wal-Mart (symbol: WMT). The gray represents the daily aggregate of the bullish posts while the black represents the aggregate of the bearish sentiment.

the predicted sentiment, but was not predictable when using the poster's own explicitly provided sentiment. Eliminating the suspicious users (i.e. the potential "pump and dumpers") to determine if predictability increased, we found no discernible difference.

References

Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? The information content of internet stock message boards. J. Financ. 59 (3), 1259–1294.

Apté, C.V., Damerau, F.J., Weiss, S.M., 1997. Data mining with decision trees and decision rules. Future Gener. Comput. Syst. 13, 197–2010.

Balakrishnan, R., Qiu, X.Y., Srinivasan, P., 2010, May. On the predictive ability of narrative disclosures in annual reports. Eur. J. Oper. Res. 202 (3), 789–901. http://www.sciencedirect.com/science/article/pii/S0377221709004822.

Ben-David, A., 2008. About the relationship between ROC curves and Cohen's kappa. Eng. Appl. Artif. Intel. 21 (6), 874–882.

Bollen, J., Mao, H., Zeng, X.J., 2011, March. Twitter mood predicts the stock market. J. Computation. Sci. 2 (1), 1–8.

Bowley, G., 2010, December. Wall St. computers read the news, and trade on it. The New York Times. http://www.nytimes.com/2010/12/23/business/23trading.html

Brown, E.D., 2012, March. Will twitter make you a better investor? a look at sentiment, user reputation and their effect on the stock market. In: Proceedings of the Southern Association for Information Systems Conference, Atlanta, Georgia.

Cao, H.H., Coval, J.D., Hirshleifer, D., 2002. Sidelined investors, trading generated news, and security returns. Rev. Financ. Stud. 15 (2), 615.

Castillo, C., Mendoza, M., Poblete, B., 2011. Information credibility on Twitter. In: Proceedings of the 20th International Conference on World Wide Web, ACM, Hyderabad, India, pp. 675–684.

Das, S.R., Chen, M.Y., 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. Manage. Sci. 53 (9), 1375–1388.

De Choudhury, M., Sundaram, H., John, A., Seligmann, D.D., 2008. Can blog communication dynamics be correlated with stock market activity? In: Proceedings of the 19th ACM Conference on Hypertext and Hypermedia, ACM, Pittsburgh, PA, pp. 55–60.

DeMarzo, P.M., Vayanos, D., Zwiebel, J., 2003. Persuasion bias, social influence, and unidimensional opinions. Q. J. Econ. 118 (3), 909.

Fawcett, T., 2004. ROC graphs: Notes and practical considerations for researchers. Mach. Learn. 31, 1–38.

Feldman, R., Rosenfeld, B., Bar-Haim, R., Fresko, M., 2011. The stock sonar – sentiment analysis of stocks based on a hybrid approach. In: 23rd IAAI Conference, San Francisco, CA.

Frieder, L., Zittrain, J., 2007. Spam works: Evidence from stock touts and corresponding market activity. Hastings Comm. & Ent. LJ. 30, 479.

Gu, B., Konana, P., Liu, A., Rajagopalan, B., Ghosh, J., 2006. Identifying information in stock message boards and its implications for stock market efficiency. In: Workshop on Information Systems and Economics, Los Angeles, CA.

Iofciu, T., Fankhauser, P., Abel, F., Bischoff, K., 2011. Identifying users across social tagging systems. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, pp. 522–525.

Ipeirotis, P.G., Provost, F., Wang, J., 2010. Quality management on Amazon Mechanical Turk. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, ACM, Washington, DC, pp. 64–67.

Jain, G., Ginwala, A., Aslandogan, Y.A., 2004. An approach to text classification using dimensionality reduction and combination of classifiers. In: Information Reuse and Integration, 2004. IRI 2004. Proceedings of the 2004 IEEE International Conference, IEEE, Las Vegas, Nevada, pp. 564–569.

Jarmulak, J., Craw, S., 1999. Genetic algorithms for feature selection and weighting. In: Proceedings of the IJCAI, Vol. 99. Citeseer, Stockholm, Sweden, pp. 28–33.

Kaymak, U., Ben-David, A., Potharst, R., 2012. The AUK: A simple alternative to the AUC. Eng. Appl. Artif. Intel. 25 (5), 1082–1089.

Kim, Y.S., Street, W.N., 2004. An intelligent system for customer targeting: A data mining approach. Decis. Support Syst. 37 (2), 215–228.

Knobbe, A., Ho, E., 2006. Pattern teams. In: Furnkranz, J., Scheffer, T., Spiliopoulou, M.

(Eds.), Knowledge Discovery in Databases: PKDD 2006. Lecture notes in computer science. Springer, Berlin, Heidelberg, pp. 577–584.

Kohavi, R., 1995. The power of decision tables. In: Lavrac, N., Wrobel, S. (Eds.), Machine Learning: ECML-95. Lecture notes in computer science. Springer, Berlin, Heidelberg, pp. 174–189.

Kohavi, R., Sommerfield, D., 1998. Targeting business users with decision table classifiers. In: Proceedings of the KDD, New York, NY. Vol. 98. pp. 249–253.

Lewis, M., 2001, February. Jonathan Lebed: Stock manipulator, S.E.C. nemesis – and 15. The New York Times. http://www.nytimes.com/2001/02/25/magazine/25STOCK-TRADER.html?pagewanted=1

Li, X., Wang, C., Dong, J., Wang, F., Deng, X., Zhu, S.,2011. Improving stock market prediction by integrating both market news and stock prices. In: Hameurlain, A., Liddle, S.W., Schewe, K-D., Zhou, X. (Eds.), Database and Expert Systems Applications. Lecture notes in computer science. Springer, Berlin, Heidelberg, pp. 279–293.

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval, first ed. Cambridge University Press, New York, NY.

Mizumoto, K., Yanagimoto, H., Yoshioka, M., 2012. Sentiment analysis of stock market news with semi-supervised learning. In: Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference, Shanghai, China, pp. 325–328.

Nonnecke, B., Preece, J., 2000. Lurker demographics: Counting the silent. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, The Hague, The Netherlands, pp. 73–80.

Ruiz, E.J., Hristidis, V., Castillo, C., Gionis, A., Jaimes, A., 2012. Correlating financial time series with micro-blogging activity. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining, ACM, Seattle, WA, pp. 513–522.

Schapire, R.E., 1990. The strength of weak learnability. Mach. Learn. 5 (2), 197–227.

Schumaker, R.P., Chen, H., 2009. A quantitative stock prediction system based on financial news. Inform. Process. Manag. 45 (5), 571–583.

Sprenger, T.O., Tumasjan, A., Sandner, P. G. and Welpe, I.M 2010. Tweets and trades: The information content of stock microblogs. SSRN 1702854. http://dx.doi.org/10.1111/j.1468-036X.2013.12007.x

Vafaie, H., De Jong, K., 1992. Genetic algorithms as a tool for feature selection in machine learning. In: Tools with Artificial Intelligence, 1992. TAI'92, Proceedings., 4th International Conference, IEEE, Arlington, Virginia, pp. 200–203.

Wang, Y.C., Joshi, M., Cohen, W., Rosé, C., 2008. Recovering implicit thread structure in newsgroup style conversations. In: Proceedings of the 2nd International Conference on Weblogs and Social Media (ICWSM II), Seattle, WA.

Wex, F., Widder, N., Liebmann, M., Neumann, D., 2013. Early warning of impending oil crises using the predictive power of online news stories. In: System Sciences (HICSS), 2013 46th Hawaii International Conference, Wailea, Maui, Hawaii, pp. 1512–1521.

Yang, Y., Liu, X., 1999. A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Berkeley, CA, pp. 42–49.

Yoon, Y., Swales, G., 1991. Predicting stock price performance: A neural network approach. In: System Sciences, 1991. Proceedings of the 24th Annual Hawaii International Conference, Vol. 4. IEEE, Kauai, Hawaii, pp. 156–162.

Yu, L., Wang, S., Lai, K.K., 2005. A novel nonlinear ensemble forecasting model incorporating GLAR and ANN for foreign exchange rates. Comput. Oper. Res. 32 (10), 2523–2541.

Zhang, X., Fuehres, H., Gloor, P.A., 2011. Predicting stock market indicators through twitter 'i hope it is not as bad as i fear'. Procedia-Soc. Behav. Sci. 26, 55–62.

Zhang, X., Fuehres, H., Gloor, P.A., 2012. Predicting asset value through twitter buzz. In: Advances in Collective Intelligence 2011. Springer, Savannah, Georgia, pp. 23–34.