Collective Sentiment Mining of Microblogs in 24-hour Stock Price Movement Prediction

Feifei Xu and Vlado Kešelj
Faculty of Computer Science, Dalhousie University, 6050 University Ave, Halifax, B3H 4R2, Canada
Email: {vlado,fxu}@cs.dal.ca

***Abstract*—We propose a method for collective sentiment analysis for stock market prediction and analyse its ability to predict the change of a stock price for the next day. The proposed method is a two-stage process, based on the latest natural language processing and machine learning algorithms. Our evaluation shows best performance with the SVM approach in sentiment detection, with accuracy rates of 71.84/74.3% for positive and negative sentiment, respectively. The results of sentiment analysis are used in predicting stock price movement (up or down), and we found that users' activity on StockTwits overnight positively correlates with stock trading on the next business day. The collective sentiments in afterhours have powerful prediction on the change of stock price for the next day in 9 out of 15 stocks studied by using the Granger Causality test.**

***Keywords*-financial forecasting; social media analytics; data mining; e-commerce;**

I. Introduction

Since the introduction of social media, companies are increasingly adopting social media technologies, using Twitter to reach out to customers or YouTube to demonstrate product features. "The wisdom of crowds," equipped with data mining rules and algorithms can automatically generate collective estimations of future performance on a variety of subjects, such as stock market performance, sports outcomes, election results and box office sales. In this work attempt has been made to study the prediction power of the collective sentiments of micro-blogging websites on the stock market.

The data used in the research are tweets collected from stocktwits.com, which is an online financial communication platform for the financial and investing community. At the time of writing, there are more than 150,000 investors on the site, which can be viewed by audiences of 40 million across the financial web and social media platforms. As a sister service to Twitter, StockTwits is composed of a large user base of trading and investing professionals, who can integrate their StockTwits accounts with their Twitter accounts if they choose to.

The objectives of this work can be presented in several steps: First, we examine performance of several natural language processing approaches to detect the public sentiment of users on StockTwits. Second, we analyze the correlation between the trade volume of stocks and the activity of users's dicussions of the stocks on StockTwits, to determine the predictive power of the social media data on the daily stock market performance.

Our ultimate goal is not to build an ideal model for stock market prediction, but to test whether the feature of social media sentiments contributes to the stock market analysis, and to assess its predictive power.

II. Related Work

The impact of news stories at the stock price has been recognized before [1]. Even though technology has emerged as one of the primary forces shaping trading markets from its inception [2], the difficulty has been the inability to capitalize on the behaviors of human traders [1]. Behavioral patterns have not been fully defined and are constantly changing, thus making accurate predictions quite difficult.

Among the most similar research to ours, done on stock market prediction from UGC (User Generated Content), is the work based on the internet forums which enable users to bet on and make market predictions about the outcomes of future events [3]–[5]. The participants include active users who constantly engage in stock market prediction, some making more accurate prediction than others; and spamming users who make random guesses. The data from the forums are fairly structured with either positive or negative votes. However, the quality of the votes cannot be traced due to the lack of reliable user profile information in the forums. Moreover, the time-sensitive nature of the financial market requires a medium where information is rapidly spreadable while forums performance is inferior in this regard. Microblogging emerges as an alternative for users to rapidly share their own opinions, and in the meantime actively follow other users opinions for both information gathering and networking purposes.

Many recent studies [6], [7] have been done on Twitter sentiments by using sentiment lexicon to simply count the positive and negative polarity words to correlate the general public sentiment to the stock market index such as Dow Jones Industrial Average (DJIA) or S&P 500. These kinds of approaches have been used as the de facto standard of much research especially financial article research primarily because of its simple nature and ease of use. However, Wilson et al. [13] wrote that the contextual polarity of the

978-1-4799-5779-8/14 $31.00 © 2014 IEEE
DOI 10.1109/CBI.2014.37

phrase in which a particular instance of a word appears may be quite different from the words prior polarity. Positive words might be used in phrases expressing negative sentiments, or vice versa. Also, quite often words that are positive or negative out of context are neutral in context, meaning they are not even being used to express a sentiment. Their work aimed to automatically distinguish between prior and contextual polarity, with a focus on understanding which features are important for this task. Because an important aspect of the problem is identifying when polar terms are being used in neutral contexts, features for distinguishing between neutral and polar instances are evaluated, as well as features for distinguishing between positive and negative contextual polarity. They evaluated how the presence of neutral instances affects the performance of features for distinguishing between positive and negative polarity. Their experiments showed that the presence of neutral instances greatly degrades the performance of these features, and that perhaps the best way to improve performance across all polarity classes is to improve the systems ability to identify when an instance is neutral.

III. Methodology and Experiment Design

The models are trained from a corpus of hand-labeled data instead of using sentiment lexicons, such as SentiWordNet for several reasons: First, although subjectivity and objectivity are taken into account in determining the tweet sentiment, in many cases objective tweets under normal circumstances should be treated as polarized tweets in this thesis. For example, the tweet "*Added to positions like $AA and $BAC...*", which will most likely be marked as objective and thus be excluded from the polarity analysis, actually implies that a user expects the stock price to rise. Secondly, the vast majority of work in sentiment analysis mainly focuses on the domains of movie reviews, product reviews and blogs, in which cases SentiWordNet represents a suitable lexicon; however, recent financial research shows that the word lists developed for other disciplines misclassify common words in financial texts [8]. For example, for the tweet "*Short $AAPL @557.50*", if regular lexicons are used, the sentiment will probably be marked as objective or neutral, while in finance the word short is a clear sign indicating that the user expects the $AAPL stock to fall.

Through the initial collection and analysis of tweets of 64 stocks, the list of stocks to be included in the analysis has been narrowed down to 16 stocks, which are the most discussed ones on StockTwits. Tweets of those stocks are collected over a period of two and a half months, along with the information of the username, date and time of publishing, source of the message, and other information which reflects the publishers profile. A complete list of the initially collected attributes can be found in Table 1.

Stock tweets sentiment detection is achieved using TagHelper tool [9], which is an application that makes use of the functionality provided by the Weka toolkit — a machine learning software written in Java and licensed under the GNU General Public License.

Table I
List of Attributes for Collection

Entity	Attributes for Collection
Tweets	Username, content of the tweets, date of publishing, time of publishing, source of the message
Users	Username, date of joining, total number of tweets by the user, number of followers, level of experience, approach of trade, holding period of trade

A. Data Acquisition and Pre-processing

Although there are fewer tweets available on StockTwits, the relatively good quality of the tweets makes it possible to take into account every tweet. Tweets of the 16 stocks, which are randomly selected and are actively discussed by StockTwits users every day, are collected for the period between March 13th, 2012 and May 25th, 2012. The period chosen did not have unusual market conditions and was a good test bed for the evaluation.

The following steps are taken to pre-process the stock tweets:
(1) Remove tweets on weekends and public holidays,
(2) Remove duplicated tweets by the same user,
(3) Replace @ sign in the tweets with text "atreplace",
(4) Replace $TICKER for each stock in the tweets with text "stocksignreplace",
(5) Replace links in the tweets with text "linkreplace", and
(6) Convert all the tweets text to be lowercase.

There are approximately 100,000 tweets for the 16 stocks included in the final analysis. In some research [10], the effect of trend was considered because the number of tweets is increasing due to Twitters increasing user base during the observation period. In this analysis, the trend effect is not taken into account since the period is less than three months. If the data is collected over a longer period, the feature values can be normalized with a time-dependent normalization factor that considers trends.

The stock data is obtained from Google Finance for the 16 stocks for the period between March 13th, 2012 and May 31st, 2012. More days are included in this set of data since the days are lagged to test the prediction power of the stock tweets. For each stock, the daily open and close price and daily traded volume are also recorded over the same period of time. The price series are then transformed into its daily relative change, i.e., if the series of price is P_i, the daily relative change will be $\frac{P_i - P_{i-1}}{P_{i-1}}$.

B. Tweet Sentiment Hand-labeling

A total number of 2380 tweets on all of the 16 stocks are hand-labeled as the training and testing data. The tweets are labeled as positive (1), negative (-1) or neutral (0). The hand-labeled data are then randomized to generate training

Table II
SOME TYPICAL TWEETS AND THEIR LABELS

Tweets	Argument	Label
The one thing I see here on $CMG is that it has touched and rallied from 401 four times since 4/24, each time making a lower high.	One can argue that it can be negative sentiment towards $CMG over the long term. However, one cannot be certain thus it is marked as neutral.	Neutral
Thinking banks will rebound soon.. $JPM	The statement is short and the positive sentiment towards $JPM is clear.	Positive
$AMZN is the paint dry yet?	"Paint dry" is a phrase which describes the process is slow and boring. The question mark reveals that the user is uncertain about it.	Neutral
$AMZN looks good to short fundamentally from tablet risk and subsidized shipping #amazon #kindle $UPS http://t.co/l2KJ3m6K	Although the word good usually represents positive sentiment, the word short is the key polarity word in this case and it is a clear sign of negative sentiment towards $AMZN.	Negative
$C 6,3% $BAC 6,26% $RIO 3,95% $DANG 7,22% $YOKU -1,41% $GOOG 2,09% $BIDU 1,73% $SOHU 8.08% $WFM 2,59% $SBUX 2,41%	In this tweet, there are lots of numbers which are the daily percentages of changes of several stocks. However, it is merely a description of the situation rather than positive or negative sentiment.	Neutral

dataset and testing dataset, which are of the amount of 2000 and 380, respectively.

For some tweets, it is difficult to label its polarity even by human efforts. In the case of vagueness, the tweets are labeled as neutral. Table 2 presents some typical tweets and their labels as well as the arguments of why they are labeled so.

The following are some general rules applied in labeling data: (1) If the tweet contains external links of long articles or numerical charts about the stocks, it is generally marked as neutral. The content of the article and the information revealed by the chart are not taken into account. (2) Positive or negative labels are only given when the sentiment can be explicitly speculated from the tweet. (3) Tweets with question marks are generally marked as neutral. (4) Simple summarizations of the stock performance by the end of the day are not taken into consideration. (5) If the user reports a loss in a subjective way instead of reporting numbers, it is fair to assume that the user has a negative feeling towards the stock; and vice versa.

C. Sentiment Detection

Models are trained and cross-validated by using the 2000 tweets, and then tested on the 380 tweets. Data is prepared before applying machine learning algorithms to train the models. The following is a description of the features used for training the models:

Punctuation: Punctuation is treated as a feature in determining whether the tweet is neutral or polarized, but not used as a feature to determine the positivity and negativity. For example, punctuation such as the question mark is a useful indicator of the uncertainty of the statement. Furthermore, the inclusion of a comma might mark that a contribution is relatively more elaborated than one without a comma.

Line length: Line length is used as an attribute because it is believed that length of contribution can sometimes serve as a proxy for depth or level of detail. In this case, lengthy contributions in tweet data contain elaborated explanations, which are important to detect in determining whether it is a neutral statement or a polarized declaration.

Unigrams and bigrams: A unigram is a single word, and a bigram is a pair of words that appear next to one another. For example, bigrams capture the contradictory meaning of the word long results between the phrases "long puts" and "long $AAPL".

Since the grammar in tweets is generally poor, part-of-speech (POS) tagging is not used in the models.

D. Approaches to Determining Collective Sentiments

By applying machine learning in NLP, each tweet is labeled with positive (1), negative (-1) or neutral (0). Multiple approaches are taken to determine the collective sentiments of each stock for each day. The following parameters, which are count of tweets, number of followers, and time of publishing, are adjusted in multiple ways to calculate the collective sentiments.

Count of tweets: If a user posted several tweets about a certain stock on the same day, those tweets are aggregated to generate the users unified sentiment, which is positive (1), negative (-1) or neutral (0). This can avoid counting the tweets with the same sentiment from the same user for multiple times.

Number of followers: In one case, the sentiments of multiple users of the same stock for the same day are added up to generate the collective sentiment; in the other case, different users unified sentiments are assigned with different weights by taking into account their numbers of followers.

Time of publishing: In light of Schumaker et al.s research [1], in which the collection period of financial articles was restricted to be between the hours of 10:30am and 3:40pm because they felt it important to reduce the impact of overnight news on stock prices, in this research similar considerations are taken by aggregating the sentiments of different periods. In one case, the sentiments of the tweets posted during 12:00am and 11:59pm on the same day are aggregated, while in the other case, the sentiments of tweets posted during the period of 4:00pm (the marketing closing time) and the next day 9:30am (the market opening time) are aggregated. It is believed that the sentiments during the open time of the market may be influenced by the real-time market fluctuations, thus create noises in the sentiments prediction power for the future. The sentiments during the

Table III
NEUTRAL VS. POLARIZED DETECTION BY SVM

Support Vector Machine	Overall Accuracy	Precision of Polarized Tweets	Recall of Polarized Tweets
Training	70.50%	70.98%	72.08%
Testing	71.84%	70.56%	71.17%

Table IV
POSITIVE VS. NEGATIVE DETECTION BY SVM

Support Vector Machine	Overall Accuracy	Precision of Positive Tweets	Recall of Positive Tweets
Training	75.68%	77.12%	82.55%
Testing	74.03%	76.23%	83.78%

closing time, however, are most likely based on the users logical and intuitive analysis of data and factual information.

E. Schema

Figure 1 represents a brief research schema of sentiment detection. As previously mentioned there is a total number of 2380 tweets on all of the 16 stocks as the training and testing data. Each tweet is equally treated by replacing the $TICKER for each stock with the identical text. The randomization process ensures that each tweet has equal chance to appear in the training data or the testing data. Out of the 2000 tweets of training data, 1028 of the tweets are labeled either positive or negative and 972 of them are labeled neutral; while out of the 380 testing data, the numbers are 181 and 199, respectively. The sentiment detection process is composed of two stages. The polarized tweets are then carried to the second stage to classify whether they are positive or negative.

Figure 2 is a description of the analytical process after the data are labeled. It mainly includes two types of analyses, which are the relationship between the stock trade volume and the tweet volume, and the relationship between stock price change and collective sentiments. As mentioned in the previous section, there are multiple approaches in determining collective sentiments. Each collective sentiment generated by different approaches is studied individually with the stock market price change. In Figure 2, GC stands for Granger Causality.

IV. RESULT ANALYSIS AND DISCUSSION

Three machine learning classifiers, Naïve Bayes, Decision Tree (J48 in Weka), and Support Vector Machine (SMO in Weka), are applied to the sentiment detection process in two stages: Neutral vs. Polarized Detection, for finding polarized tweets, and then Positive vs. Negative Detection, which is applied only on polarized tweets. The overall accuracy rates of the best classifiers at the two stages are 71.84% and 74.3%, while the numbers for ZeroR baseline (simple majority classification) models are merely 51.4% and 57.98%, respectively. These are satisfactory results, as the process of determining the sentiment of a tweet is not very consistent even among humans. It has been shown that and people only agree on sentiment 80% of the time [11].

A. Analysis of Predictive Power of Social Media

Volume Correlation: Bivariate correlation analyses are conducted to study the correlations between the unified volume of tweets and the trading volume. The Pearson product-moment correlation coefficient is used as a measure of the strength of linear dependence between two variables.

Table V
VOLUME CORRELATION FOR 12:00AM–11:59PM

$AAPL	$AMZN	$BAC	$BIDU	$C	$CMG
0.871	0.855	0.936	0.858	0.792	0.830
$FSLRL	**$GOOG**	**$GS**	**$IBM**	**$JPM**	**$MSFT**
0.868	0.829	0.73	0.591	0.834	0.543
$NFLX	**$PCLN**	**$RIMM**	**$YHOO**		
0.851	0.729	0.957	0.390		

Table VI
VOLUME CORRELATION FOR 4:00PM–9:30AM

$AAPL	$AMZN	$BAC	$BIDU	$C	$CMG
0.530	0.566	0.602	0.757	0.654	0.745
$FSLRL	**$GOOG**	**$GS**	**$IBM**	**$JPM**	**$MSFT**
0.760	0.454	0.540	0.679	0.772	0.681
$NFLX	**$PCLN**	**$RIMM**	**$YHOO**		
0.677	0.226	0.359	0.503		

From Table V, we see that there are significant positive correlations (with critical value of approximately 0.3) on the same day for all of the 16 stocks for the period of 12:00am–11:59pm. For the time period of 4:00pm–9:30am in Table VI, the significant positive correlations (with critical value of approximately 0.3) are found in 15 out of the 16 stocks. Although the time stamps of the two time series are lined up for the period of 4:00pm–9:30am, the period during which the tweets are published is actually before the daily trading starts. Thus there is certain prediction power that can be seen from this correlation. It is valuable information that users activity on StockTwits overnight significantly positively correlates to the stock trading volume the next business day. This can be explained by the relationship between the market capitalization and the daily trading volume, and the relationship between the daily trading volume and the public attention. Although it is not likely to be causality, it is still clear evidence that the user base on StockTwits is a decent representation of the public who engages in stock market.

Sentiment and the Stock Market: Granger Causality test is carried out to study whether prediction power exists between users collective sentiments of tweets and the stock market price. Each pair of the series of the 16 stocks is studied individually. It is not to test the actual causation, but whether one time series has predictive information about the other. Granger Causality has been widely used in economics

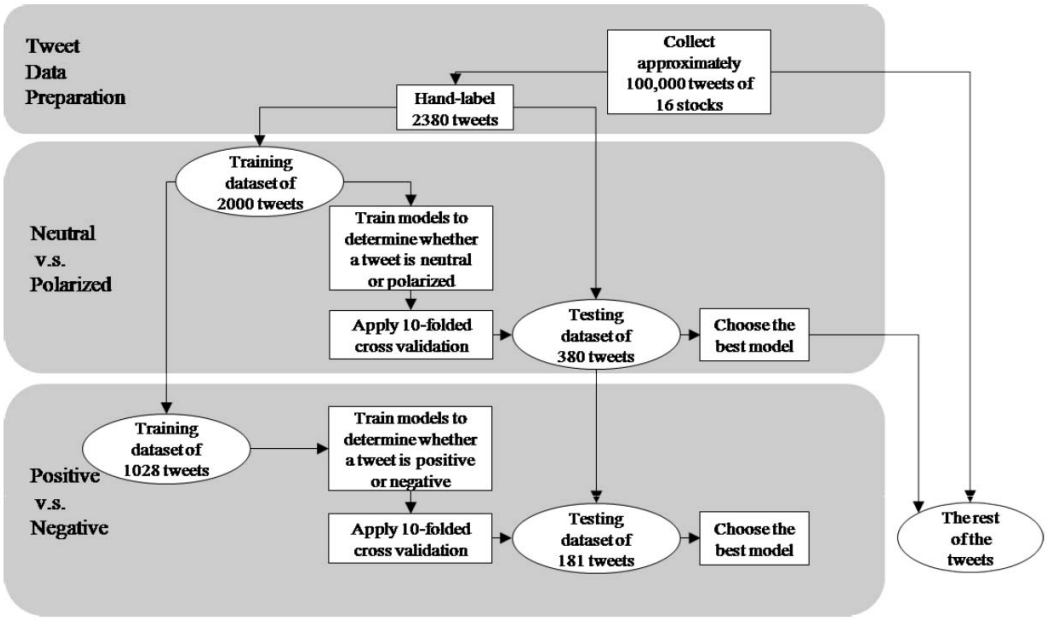

Figure 1. Research Schema of Sentiment Detection

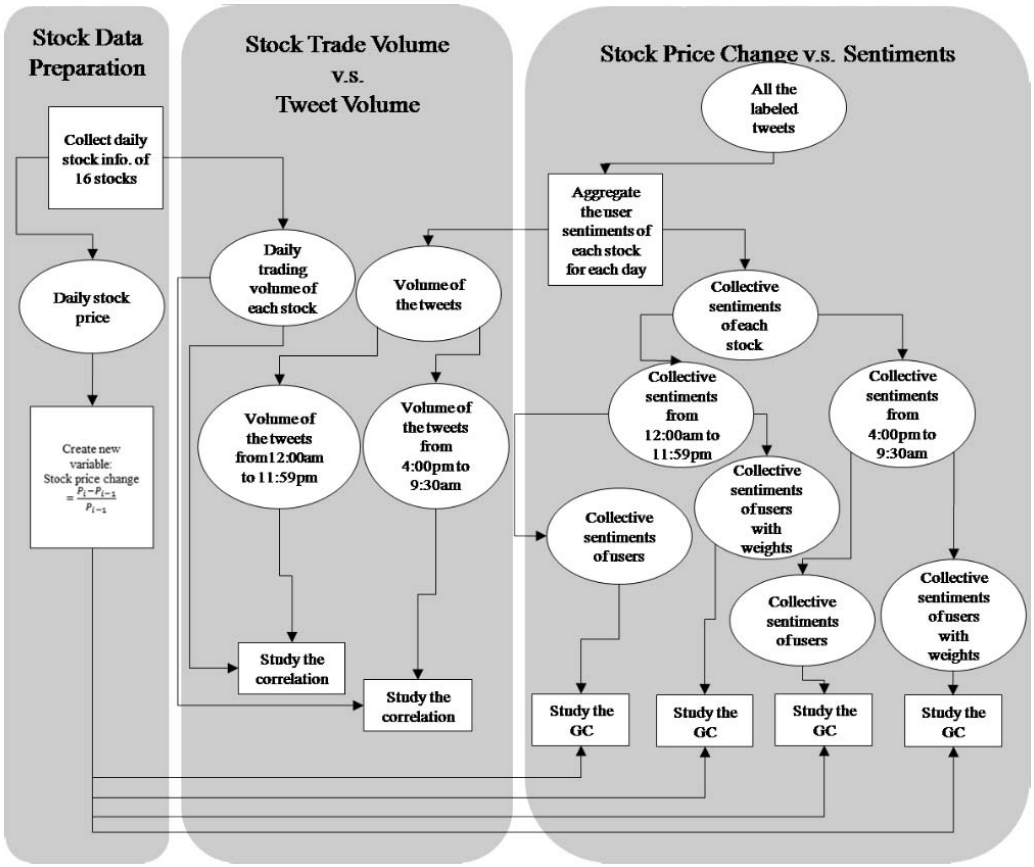

Figure 2. Research Schema of Data Analysis

since 1960s. Its mathematical formulation is based on linear regression modeling of stochastic processes [12]. The basic GC definition is quite simple. Suppose that there are two terms X_t and Y_t, and the first attempt is made to forecast X_{t+1} using past term X_t, and then the second attempt is made to forecast X_{t+1} using past terms of X_t and Y_t. If the second forecast is found to be more successful, according to standard cost functions, then the past of Y appears to contain information helping in forecasting X_{t+1} that is not in past of X. In particular, there could be other possible explanatory variables. Thus, Y_t would "Granger cause" X_t if (a) Y_t occurs before X_{t+1}; and (b) it contains information useful in forecasting X_{t+1} that is not found in a group of other appropriate variables. In the case of this work, if Y is the collective sentiment and X is the stock price change, a GC test can demonstrate whether the collective sentiment appears to contain information helping in forecasting the stock price change of tomorrow that is not in the stock price change of today.

To apply Granger Causality test, each time series needs to be stationary. Although Bollen et al.s method of stationarizing the series is a good approach, which is to normalize the series to z-scores on the basis of a local mean and standard deviation within a sliding window of a certain number of days before and after the particular date [6], it can only be used to study the causality but does not have practical application because one can never get data from the future to study the present. For that reason, Augmented Dickey Fuller (ADF) Unit Root Test of each series is examined individually to check its stationarity before the GC test. Normally, if the series cannot pass the ADF test, the first difference should be taken to stationarize the series. However, there is not much rational reasoning of taking the first difference in this dataset both due to its coarse nature (it is estimated from the social media) and the assumption that the day-to-day collective sentiments on the stock market should be random. If the series fails to pass the ADF test, it will be dropped from the analyses.

Six different types of collective sentiments are calculated and are tested on its stationarity by using ADF as described on all of the 16 stocks. Out of a total of 96 series, majority of the series passed the ADF test with 10 exemptions (greyed in Table VII).

The test is conducted at the lag of 1 (p is 1) because the volume correlation reveals that users activity on StockTwits overnight significantly positively correlates to the stock trading volume the next business day, so the relationship may also exist for the collective sentiments and the stock price change. A total number of 86 one-to-one Granger Causality tests are systematically carried out to study the relationship between the collective sentiments and the change of stock. The following Table : Results of Granger Causality Table VII presents a summarization of the Granger Causality test results. Graphs of some series are also analyzed for further information.

The "Afterhours Simple Sum", which is the simply summed up collective sentiments for afterhours, has powerful prediction on the change of stock price for the next day in most of the stocks studied. For "Afterhours Simple Sum", collective sentiments G-cause change of stock in 9 out of 15 tested stocks; and change of stock G-causes collective sentiments in 4 out of 15 stocks. When the same collective sentiments are normalized by the numbers of users, however, the prediction power becomes dissolved by the noises generated by having too many users count into the equation. Similarly, when the unified sentiments are assigned with weights, which are the numbers of the publishers followers, the prediction power is not as strong as simply summed up sentiments. On the other hand, the reverse process that changes of stock G-causes collective sentiments is also found in several stocks by taking different approaches. Interestingly, the one with the most relationships is again the "Afterhours Simple Sum" approach of generating collective sentiments; and 3 out the 4 relationships are in completely different stocks. It is profound finding that the Granger Causality exists in different directions in different stocks. For some stocks, the collective sentiments have prediction power on the stock price change for the next day; while for some other stocks, the stock price change actually influences users' collective sentiments for the next day.

Sentiment and the Direction of the Stock Movement: If "Afterhours Simple Sum" collective sentiments are used to predict the bi-directional stock movement for the next day without considering the magnitude of the movement, the co-movements of the two series can be noticed in most of the series. For example, the bi-directional co-movement can be seen 72.5% times on the $RIMM time series, and 70% of times on the $YHOO time series. Figure 3 presents the scatter plot of bi-directional stock movements by using "Afterhours Simple Sum" sentiments and the afterhours unified volumes of tweets before the market opens. It seems that the two variables are better in predicting negative stock movements. When the sentiment is negative, most of the stock movements are negative. When the sentiment is negative and the volume is large, the prediction power of negative stock movements becomes even stronger. Similar observations can be found for positive stock movements when the sentiment is positive and the volume is large. However, the data is too scarce to draw a conclusion.

V. Conclusion and Future Work

The stock market prediction has been intensively studied by scholars and professionals from finance domain, information management domain and even psychology domain. Social interaction is an important aspect of the decision-making process, and the use of online social media, a novel form of communication emerged over the past five years,

Table VII
RESULTS OF GRANGER CAUSALITY TEST

Stock	Full Day Simple Sum	Afterhours Simple Sum	Full Day Simple Sum Normalized	Afterhours Simple Sum Normalized	Full Day Weighed Sum	Afterhours Weighed Sum
$AAPL			x	x	x	x
$AMZN		$S \xrightarrow{at\ 0.05} C$			x	x
$BAC		$S \xrightarrow{at\ 0.05} C$ $C \xrightarrow{at\ 0.05} S$	x	x	$S \xrightarrow{at\ 0.10} C$	$C \xrightarrow{at\ 0.05} S$
$BIDU		x	x	x	$S \xrightarrow{at\ 0.05} C$	x
$C	x	x	x	x	$S \xrightarrow{at\ 0.05} C$	x
$CMG	x $C \xrightarrow{at\ 0.05} S$	$S \xrightarrow{at\ 0.05} C$	x $C \xrightarrow{at\ 0.05} S$	x	x	x
$FSLR	x	$S \xrightarrow{at\ 0.05} C$	x	x	x	$S \xrightarrow{at\ 0.05} C$
$GOOG		x		x	x	x
$GS	x $C \xrightarrow{at\ 0.1} S$	x $C \xrightarrow{at\ 0.1} S$	x	x	x $C \xrightarrow{at\ 0.05} S$	x $C \xrightarrow{at\ 0.05} S$
$IBM	$S \xrightarrow{at\ 0.05} C$	$S \xrightarrow{at\ 0.05} C$	x	x	x	x
$JPM	x $C \xrightarrow{at\ 0.05} S$	x $C \xrightarrow{at\ 0.05} S$	x		x	x
$MSFT	$S \xrightarrow{at\ 0.05} C$	$S \xrightarrow{at\ 0.05} C$	x	x	x	x
$NFLX	x	$S \xrightarrow{at\ 0.05} C$	x	$C \xrightarrow{at\ 0.1} S$	x	$S \xrightarrow{at\ 0.05} C$
$PCLN	x	x $C \xrightarrow{at\ 0.1} S$	x	x	x	x
$RIMM	x	$S \xrightarrow{at\ 0.05} C$	x	x	x	x
$YHOO	x	$S \xrightarrow{at\ 0.05} C$	x	x	x	$S \xrightarrow{at\ 0.05} C$ $C \xrightarrow{at\ 0.05} S$

Notation: Full Day: 12:00am–11:59pm; Afterhours: 4:00pm–9:30am; Simple Sum: Simply summed up collective sentiments; Weighed Sume: Weighed summed up collective sentiments; Simple Sum Normalzied: Simply summed up collective sentiments divided by the unified volume of tweets during the same period; x: No signfificant relationship; C: Change of stock; S: Sentiment; at 0.05, at 0.1: Probability of accepting the Null Hypothesis.

has started gaining enough data for carrying out analysis. This work has empirical contributions to this research area by taking a unique approach of public sentiment analysis and also has practical implications by proposing an effective technique of stock market analysis.

The research architecture consists of a NLP approach and a statistical analysis approach. The NLP approach of sentiment detection is again a two-stage process by implementing Neutral v.s. Polarized detection before Positive v.s. Negative detection. The two-stage approach is in line with Wilson et al.s research [13] on how the presence of neutral instances may affect the performance of features for distinguishing between positive and negative polarity. The statistical approach takes a unique path in dealing with the time in light of the thought that the sentiments during the open time of the market may be influenced by the real-time market fluctuations thus create noises in the sentiments prediction power for the future. An important conclusion is that the collective sentiments of afterhours are much stronger predictors.

Our initial assumption that the users on StockTwits have the genuine incentive to produce high-quality content is validated through the hand-labeling process as well as the proven prediction power. Besides, attempt has also been made on weighing the sentiments of different users by putting into the equation the measurements of experts followers to reflect their public influence. However, the prediction power is not as strong as simply summed up sentiments. This could be due to the over-simplified method of weighing the unified sentiments with the number of followers.

Some of the practical implications of this work follow from having:
(1) demonstrated that Support Vector Machine is the best classifier with the overall accuracy rates of 71.84% and 74.3%, respectively, at the two-stage sentiment detection process;

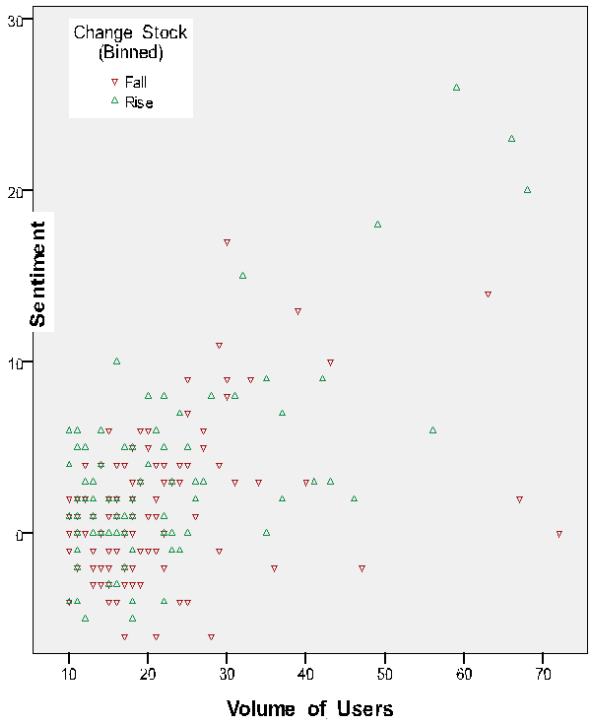

Figure 3. Scatter plot of Bi-directional Stock Prediction

(2) discovered that users activity on StockTwits overnight significantly positively correlates to the stock trading volume the next business day;
(3) determined that simply summed up collective sentiments for afterhours has powerful prediction on the change of stock price for the next day in 9/15 of the stocks studied by using Granger Causality test; and
(4) discovered that the overall accuracy rate of predicting the up and down movement of stocks by using the collective sentiments is 58.9%.

The overall accuracy rates of 71.84% and 74.3% are satisfactory results, as the process of determining the sentiment of a tweet is vague even for human and people only agree on sentiment 80% of the time. The fact that users activity on StockTwits overnight significantly positively correlates to the stock trading volume the next business day is clear evidence that the user base on StockTwits is a decent representation of the public who engages in stock market. It can also be concluded from the analyses that the collective sentiments of afterhours have certain prediction power on the direction of the stock movement of the next day.

For future work, our plans include analyzing data over a longer period, using an expanded lexicon, and use of user profile features.

ACKNOWLEDGMENT

The authors would like to thank Dr. Vladimir Lučić from Barclays Capital for discussions and comments on the work, and to acknowledge support from the NSERC CRD grant program.

REFERENCES

[1] R. Schumaker, Y. Zhang, and C. Huang, "Evaluating sentiment in financial news articles," *Decision Support Systems*, vol. 53, no. 3, pp. 458–464, 2012.

[2] R. T. Williams, *An Introduction to Trading in the Financial Markets: Trading, Markets, Instruments, and Processes: Trading, Markets, Instruments, and Processes*. Academic Press, 2011.

[3] B. Gu, P. Konana, A. Liu, B. Rajagopalan, and J. Ghosh, "Predictive value of stock message board sentiments," *McCombs Research Paper No. IROM-11-06*, 2006.

[4] S. Hill and N. Ready-Campbell, "Expert stock picker: the wisdom of (experts in) crowds," *International Journal of Electronic Commerce*, vol. 15, no. 3, pp. 73–102, 2011.

[5] C. Avery, J. A. Chevalier, and R. J. Zeckhauser, "The "caps" prediction system and stock market returns," National Bureau of Economic Research, Tech. Rep., 2011.

[6] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[7] H. Pajupuu, R. Altrov, and K. Kerge, "Lexicon-based detection of emotion in different types of texts: Preliminary remarks," *Eesti Rakenduslingvistika Ühingu aastaraamat*, no. 8, pp. 171–184, 2012.

[8] T. Loughran and B. McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.

[9] C. Rosé, Y.-C. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger, and F. Fischer, "Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning," *International journal of computer-supported collaborative learning*, vol. 3, no. 3, pp. 237–271, 2008.

[10] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes, "Correlating financial time series with micro-blogging activity," in *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012, pp. 513–522.

[11] S. Grimes, "Expert analysis: Is sentiment analysis an 80% solution?" *InformationWeek (March 2010). Available online at http://www. informationweek. com/news/software/bi/showArticle. jhtm l*, 2010.

[12] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[13] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis," *Computational linguistics*, vol. 35, no. 3, pp. 399–433, 2009.