Social Data Analysis for Stock Market Indicator

Chanon Hongsirikulkit  
University of Illinois at Urbana-Champaign  
Email: hongsir2@illinois.edu

Hathfi B. Hakim  
University of Illinois at Urbana-Champaign  
Email: hhakim2@illinois.edu

Abstract—In the early day of finance field, there is a theory that stock prices are driven by only one factor, financial factor, such as company income, net profit, or expense. However, the modern theory which contrasts with the older one states that emotion also plays an important role in the stock market movements, people make a decision to buy or sell stock based on their emotions. In order to predict the stock market trends, we have to know overall people’s sentiments at that time also. The question is how can we get those sentiments. Today social network becomes very popular, people around the world share about everything in their daily life, such as food, work, travel, politics, and many others. This kind of data is generated in every second making the social network are richer with knowledge and also reflect the overall sentiment of people at a specific time. Many research papers set up the experiment to exploit the benefit from this unstructured data source. They proposed various methods that analyze mood of people on the social networks, such as Twitter and Yahoo Finance Board, to predict the stock market movements. Our research will show that sentiment information from social network can indicate stock price movement, and topics in each day can explain that correlation.

Index Terms—Social Data, Twitter, Sentiment Analysis, Topic Modeling, Stock Market Indicator

I. INTRODUCTION

Today social network becomes very popular, people share news, ideas, and opinions about everything on this world-wide communication platform. This platform stores hidden valuable knowledge which can be discovered by using appropriate methodologies. Many studies propose the way to gain the knowledge from the social network platform, each approach is different in terms of performance and application to be applied. For example, [2] uses tweet-rate and Twitter sentiment to predict box-office revenues for movies. They found strong correlation between information on Twitter and box-office revenues, and their prediction model also works well compared to the golden standard Hollywood Stock Exchange. Another example [10], the authors detect the tweets which contain information needs and analyze their characteristics. The results show that the information needs, question tweets, are different from general posts, they convey information that reflect the real-world demands. We can apply this knowledge to improve actual applications, such as search engine system. For our research, we will focus on using social network as a stock market movement indicator which will be described in following sections.

A. Twitter Cashtag

Our research project is focus on analyzing Twitter message data which is called tweet as an input for the stock market prediction. One of the advantages of this tweet data is the existence of “Cashtags”: “Cashtag” are stock ticker symbols that are prefixed with a dollar sign. For example, to tweet about Twitter stock, users usually would use $TWTR; for Amazon and Facebook, they would use $AMZN and $FB. Following the same strategy as with hashtags, Twitter made cashtags clickable in July 2012 [5]. A click on a cashtag results in a search for tweets containing this cashtag. By filtering the tweets using this cashtag property, it will reduce the noise data which are not relevant for our research purpose to indicate the stock market price.

B. Twitter API

Another reason using this social media platform is because it provides a robust API which enables user to collect the dataset. There are two ways to collect tweets from the Twitter. The first one is from the Twitter Searching API 1. This only allows searching tweets from one week in the past for free. However, the Search API only return a sample dataset of the matched searched tweet criteria. The other way is using Twitter Streaming API 2. It allows collecting the real time tweets rather than search from the history. A tweet downloaded from Twitters API contains a series of fields in addition to the text of the message in a JSON 3 format. For example, each JSON record includes meta-data such as posting date, and information about

1 https://api.twitter.com/1.1/search/tweets.json
2 https://stream.twitter.com/1.1/statuses/filter.json
3 http://www.json.org/
its author at the time of posting such as the number of followers, the date of account creation, etc. We consider that this meta-data would be very useful for our research.

II. RELATED WORK

A. Sentiment Analysis

In the early day, the efficient market hypothesis (EMH) states that the movement of stock price is based on rationale and available fundamental information only. However, the behavioral finance argues that investors behavior can also drive the stock market movement. Investors tend to use their biases and emotions to make decision to buy or sell stocks. [3] sets up the experiment to verify this hypothesis.

The researchers crawl data from Twitter about 9 million tweets from February 28 to December 19, 2008. They use sentiment analysis tools, GPOMS and OpinionFinder (OF), to extract people’s emotions on Twitter. In the paper, they examine data in the two-month period from October 5, 2008 to December 5, 2008 which has some important events that can be captured, presidential election and Thanksgiving.

To validate data, the authors plot sentiment scores extracted from GPOS and OF in the time series, and they found that people have anxiety on the day before election, since the calm score significantly decreases and have sure feeling after election day based on GPOS results, and the OF also gives the high positive feeling score. On Thanksgiving day, the happy and positive scores seem to be high which also reflect the real-world phenomenon.

After data validation, the authors verify the hypothesis that emotions play an important role in stock market by using Granger causality analysis to test the correlation between sentiments and DJIA values and applying self-organizing fuzzy neural network (SOFNN) to measure the forecasting capability of sentiment on stock market prices. The results show that the calm value is the best indicator for stock market movement which has time lags about 2 - 6 days. This paper can be considered as a fundamental research of using sentiment on social network to predict stock price. Many recent researches develop the new methodology and prediction model based on the principles in this paper.

Many researches occur to verify behavioral finance concept and exploit the knowledge on social network for stock market prediction. Some of them use only sentiment to forecast stock market movement, while others use both topic and sentiment to enhance the prediction capability. We can describe some related researches in this field as follows.

B. Apply sentiment to predict stock market movement

We would like to introduce some related works using sentiment from social network to predict stock movement. This knowledge can be applied to various applications. Some examples are described as follows. [1] sets up the experiment to show that sentiment data from Twitter can predict individual stock market prices and also movie box-office revenues. For stock market application, they focus on technology company stocks, namely, Apple (AAPL), Google (GOOG), Yahoo! (YHOO), and Microsoft (MSFT). For box-office revenue application, they use Green Lantern to be a case study. The authors propose various forecasting models, each of them is based on different types of machine learning classifier, including support vector machines (SVM) with different kernel types, various degrees of neural networks, and linear models, and then compare the results to find the highest accuracy model. They also introduce the sentiment measure, called Sentiment Index, in the paper.

A recent work [9] finds the correlation between sentiment and volume on Twitter with stock price returns. The authors gather tweets via Twitter Search API focusing on the posts which contain stock cash-tag. They construct the classifiers for sentiment analysis, extracting features, such as unigrams and bigrams, and using experts to annotate tweets. With support vector machine classifier, they can automatically classify all tweets into 3 categories: positive, neutral, and negative.

To validate the forecasting capability of Twitter sentiment, they introduce sentiment value, called Sentiment Polarity, and then use Pearson Correlation and Granger causality to test their data.

The results show that there is small correlation between sentiment information and stock price returns because there is a more complex relationship between these two dimensions. To detect this correlation, they propose new approach, named event study, which is similar to the technique widely used in finance field. The general idea is the authors detect events on Twitter by using tweet volume, identify the events that can affect stock price returns by applying sentiment polarity, and then compare the results with the ones given by a financial source. The interesting knowledge that they found during the experiment is some events detected by their approach, but not by financial technique, cause impact on stock price returns, and there is a strong correlation between sentiment information on social network and stock market movement.

C. Apply sentiment and topic to predict stock market movement

Some related works [1] use only sentiment information as a feature for forecasting stock price. However,
some approaches [8] exploit topic modeling technique to enhance the capability of prediction model. They show that combining topic and sentiment information can increase the accuracy of their model compared to model with topic or sentiment analysis only.

This paper proposes a new approach for extracting both sentiment score and topic from the social network, called Aspect-based sentiment, and use them as features for a classifier. Many previous works test their prediction model on a small size dataset, about 15 transaction dates, but this new method can perform on one-year period dataset effectively. The authors choose Yahoo Finance as an information source.

For feature extraction, they use data from two sources, the historical stock prices of 18 stocks from Yahoo Finance and the sentiment data from Yahoo Finance Message Board. Based on support vector machine (SVM) classifier, the authors construct various models with different features to verify that their approach giving the highest accuracy.

Finally, they evaluate the accuracy of each prediction model and found that the Aspect-based method can predict stock movement correctly up to 54.41% of test dataset which is higher than other models. This experiment shows that topic modeling can be applied to increase accuracy of prediction model.

To summarize, sentiment information on social network are effective in terms of forecasting capability. Many researches verify this hypothesis by conducting experiments to analyze the correlation between sentiment data and real-world events. They found both direct and indirect relationships between people’s emotion and stock market movement, and applied these knowledge to construct prediction model. Moreover, combining sentiment information with topic is another approach which can enhance accuracy of forecasting model, this approach also performs well on large dataset.

III. METHODS

A. Dataset

In this study, we collect twitter data through the Twitter streaming API by using 3 most popular cashtag symbol, which is $AAPL$ for Apple, $GOOG$ for Google, and $YHOO$ for Yahoo. We run python script for 24 hours a day on a server for one month from April 1 to April 29, 2016. The Nasdaq Stock market in which the three companies was registered are open on Monday to Friday and closed on weekends, so on that data collecting period, we get the twitter data for about four-week of stock market trading session. The python script writes each tweet resulting from Twitter streaming API on a file with JSON format. We separate the twitter data for each day in each different file.

Once we get all the twitter data file, we perform a data cleansing to remove tweets which use language other than English. This is a very easy task to do on twitter dataset because the data contains metadata that informs the language which is “{'lang': 'en’}”

During the period, we collected about 137,560 tweets for cashtag $AAPL$, 51,060 tweets for cashtag $GOOG$ and 17,360 tweets for cashtag $YHOO$ with total of 205,980 tweets. The distribution of collected tweets of each day in April 2016 is shown in the figure 1. In that graphs, it can also be seen that there is a tendency of increasing number of tweets on Monday to Friday and a relative declining on Saturday and Sunday. At a certain period, also seen the number of tweets rose sharply, this is usually happen because there is an event that occurred on the day.

B. Sentiment Analysis

For sentiment analysis part, we construct the classifier for sentiment classification. There are three features considered for labeling tweets which are described as follows.

1) POS Tagging - Although we filter only relevant tweets based on cashtag, the messages in tweets still contain noise. Some words, such as specific name or link, are not useful for sentiment classification. We decide to apply NLP technique, called POS Tagging to assign parts of speech to each word, such as noun, verb, adjective, and adverb, and then filter only adjective words as features. We assume that adjective word should represent sentiment of message better than other types of word. This technique is performed by using NLTK software.

2) Unigram - Using unigram seems work well in many previous researches. We tokenize words in tweets and filter only adjective words. Then, we select only top 100 frequent words as features. However, we also try to apply bigram with tweet messages, but the result is not meaningful. Most of frequent bigrams contain noise words which do not convey any sentiment information, so we decide to use only unigram feature for the classifier.

3) Word Lexicon - We use opinion words from [6] and [7] as features. These words are categorized into two groups, negative and positive words. The unigrams extracted from training data are not sufficient to classify sentiment of tweets, some opinion words occur in tweets can not be found in training data; therefore, the word lexicon is used to close this gap.

After feature extraction, we construct the classifier by using 5 weak classifiers, namely, Naive Bayes, Multi-
nominal Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, and Linear Support Vector, to do majority voting. Then, we train the classifier with 10000 short product reviews data, 5000 positive and 5000 negative reviews and test with 2000 reviews. The accuracy of classifiers are shown in Table I.

Finally, we apply the classifier to label all tweets and select only labeled tweets which have 100% confidence, all 5 classifiers give the same label, to calculate sentiment polarity for each day. Sentiment polarity is calculated by using equation as in Eq. 1

$$\text{Sentiment polarity} = \frac{p - n}{p + n}$$  \hspace{1cm} (1)

- $p$: numbers of positive tweets
- $n$: numbers of negative tweets

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>68.95</td>
</tr>
<tr>
<td>Multinomial Naive Bayes</td>
<td>65.90</td>
</tr>
<tr>
<td>Bernoulli Naive Bayes</td>
<td>64.60</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>63.40</td>
</tr>
<tr>
<td>Linear Support Vector</td>
<td>63.55</td>
</tr>
</tbody>
</table>

TABLE I: sentiment classifier accuracy on short product reviews

C. Topic Mining

We use ToPMine [4] to do topic mining and also construct good phrases from tweet data. The general idea of ToPMine algorithm is to perform phrase construction before topic modeling. First, this model will use frequent contiguous pattern mining to extract...
candidate phrases and their counts from tweet collection. Then, using agglomerative merging of adjacent unigrams based on significance score to segment documents into "bag of phrases". Finally, the model will apply PhraseLDA to segmented phrases with the constraint that all words in a phrase have to be in the same topic. The results from this algorithm are topics with their distributions and good phrases extracted from tweets.

According to our tweet data, we try to do topic mining in daily and monthly periods. However, the results show that there are not many topics occur in one day or one month, and they do not convey sufficient information for describing the correlation between sentiment and stock price. Nevertheless, good phrases generated by ToPMine seems to be useful and can explain some correlations occurring in the time series. Therefore, we decide to use top 30 frequent phrases extracted from tweets in each day to represent topics in that day.

Our algorithm used to extract topics in each day is described as follows.

1) Use ToPMine to construct good phrases from tweets in one month.
2) Read tweets in each day and count occurrences of phrases derived from step 1.
3) Select only top 30 frequent phrases for representing the topics in each day.
TABLE II: topic phrases on April 7 and April 27

<table>
<thead>
<tr>
<th>April 7</th>
<th>April 27</th>
</tr>
</thead>
<tbody>
<tr>
<td>iPhone SE</td>
<td>Huge Wins</td>
</tr>
<tr>
<td>Apple Watch</td>
<td>Waiting for the Fed</td>
</tr>
<tr>
<td>Early movers</td>
<td>Apple Gives Us Huge</td>
</tr>
<tr>
<td>Tim Cook</td>
<td>billion in cash</td>
</tr>
<tr>
<td>Apple CEO</td>
<td>billion left</td>
</tr>
<tr>
<td>CEO Tim Cook</td>
<td>market cap</td>
</tr>
<tr>
<td>Tim Cook is joining</td>
<td>Bears today</td>
</tr>
<tr>
<td>Tim Cook is joining</td>
<td>iPhone sales</td>
</tr>
<tr>
<td>joining the board</td>
<td>NHL teams</td>
</tr>
<tr>
<td>human rights</td>
<td>NFL teams</td>
</tr>
<tr>
<td>Apple CEO Tim Cook</td>
<td>NBA teams</td>
</tr>
<tr>
<td>HFT Algos</td>
<td>MLB teams</td>
</tr>
<tr>
<td>iPhone hack</td>
<td>Tim Cook</td>
</tr>
<tr>
<td>tech you need to know today</td>
<td>Wall St</td>
</tr>
<tr>
<td>bad for Apple</td>
<td>earnings were awful</td>
</tr>
<tr>
<td>things in tech</td>
<td>Real Madrid</td>
</tr>
<tr>
<td>things in tech you need to know today</td>
<td>st time</td>
</tr>
<tr>
<td>FBI says it only knows how to hack</td>
<td>Wall Street</td>
</tr>
<tr>
<td>key Apple</td>
<td>market cap shrinks</td>
</tr>
<tr>
<td>Apple Watch Herm</td>
<td>earnings decline</td>
</tr>
<tr>
<td>Apple iPhone</td>
<td>Dead body</td>
</tr>
<tr>
<td>Apple FBI</td>
<td>Jay Z And Beyonce</td>
</tr>
<tr>
<td>San Bernardino</td>
<td>rumors about the iPhone</td>
</tr>
<tr>
<td>sharing iPhone</td>
<td>Increased to Buy</td>
</tr>
<tr>
<td>San Bernardino iPhone</td>
<td>Rating Increased</td>
</tr>
<tr>
<td>Price Target</td>
<td>Apple shares</td>
</tr>
<tr>
<td>Apple iPhone SE</td>
<td>things in advertising you need to know today</td>
</tr>
<tr>
<td>Wall St</td>
<td>weak earnings</td>
</tr>
<tr>
<td>Wall Street</td>
<td>conference room</td>
</tr>
<tr>
<td>Cookie Monster</td>
<td>Skip Bayless</td>
</tr>
</tbody>
</table>

IV. EXPERIMENT AND RESULT

A. Sentiment Analysis

During the experiment we found some quite interesting results as shown in figure 2, figure 3, and figure 4. The upper bar chart of each figure is the movement of the stock price of each day in April 2016. The blue bar chart indicates that the stock price is going up and the red bar indicates that the stock price is going down. While the lower bar chart is the twitter sentiment polarity as described in section III. Following the upper bar chart, the blue bar chart in the lower chart indicates that the sentiment is positive and the red bar chart indicates that the sentiment is negative.

Figure 2 shows the comparison between the movement of the stock price of AAPL (Apple Inc) and the tweets sentiment polarity filtered by $AAPL$ cashtag. For this dataset, we get accuracy about 62%. 13 out of 21 days of observation match between the sentiment polarity and the stock price movements. Similarly, figure 3 presents the observation for Google Inc stock price, the hash tag used for this is $GOOG$. This time, the charts match for the first six days in a row. Overall, 15 out of 21 days of observation is matched, therefore, the accuracy is about 71%.

We also perform the sentiment analysis for Yahoo Inc stock price with the $YHOO$ cashtag as shown in figure 4. The result is slightly worse than the other company but overall the accuracy is still acceptable (about 52%). For this $YHOO$ cashtag, we collect so many tweets from the Twitter Streaming API. However, those tweets do not only contain the $YHOO$ cashtag but all tweets related with “yahoo”. So we did a cleansing process and use only the tweets which contain $YHOO$ cashtag.

To improve the accuracy of sentiment classifier, we try to use training data which contain tweets about specific stocks. We combine tweets from Sentiment 140\(^4\) and Sentiment Analysis Dataset\(^5\), select the ones which contain symbol "AAPL", "GOOG", and "YHOO". Totally, we have about 4500 tweets. Then, merging this additional data to the original short product reviews to create new training set. The results show that the accuracy does not significantly change compared to the current one. There are two assumptions that can explain this outcomes. First, the additional dataset might be too small, so it does not affect the overall accuracy of classifier. Second, the tweets data are very sparse; therefore, the training set does not convey discriminative features for classifying the real dataset.

\(^4\)http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip
B. Topic Mining

We analyze topic information to see whether it can describe the correlation between sentiment information and stock price. For positive correlation case, the top 30 phrases do not show the positive sign about individual stock price obviously. However, for negative correlation case, the phrases can explain the correlation better.

For example, on April 7, we can see phrases, such as "iPhone hack", "bad for Apple", and "San Bernardino". The news about San Bernardino terrorist is related to Apple directly and causes the decreasing of stock price in that day. Another example on April 27, some negative phrases, such as "earnings were awful", "market cap shrinks", and "earnings decline"; these are all bad signs for Apple stock resulting in significantly dropping of AAPL price. All topic phrases on April 7 and April 27 are shown in table II.

V. CONCLUSION

As a popular micro-blogging platform, people use Twitter not only to express their feelings but also to respond to a case or event into a status or tweet. The opinion of people in their tweet contains knowledge which has attracted so many researchers to study this dataset. Supported by the robust API, researchers could collect so many data to study for a different purpose.

In this study, We utilized the cashtag property of this platform. People share their opinion about specific stock market symbol using this property.

In this study, we show that people tend to tweet using cashtag more in weekdays and less in the weekend which is following the open time of the stock market. There is also a significantly increasing number of tweets when some event happens that can affect a specific stock price of a company.

Sentiment information is a one of good indicators for stock price movement. Overall, sentiment from Twitter can indicate stock price with accuracy more than 50%. The accuracy of the individual stock is vary depending on factors which affect the stock at that time. Sentiment information is just one factor reflecting the stock price, there are other factors which indicate the stock movement, such as fundamental financial information of company and events that cannot be captured by the social network.

Topic is another indicator which is useful for stock market analysis. It can explain the correlation between sentiment information and stock price movement. According to our experiment, the topic in each day can describe negative correlation better than positive correlation. It contains more informative data which show explicit bad sign about the stock price.

REFERENCES