Expert Systems With Applications 42 (2015) 9603–9611

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Sentiment analysis on social media for stock movement prediction

Thien Hai Nguyen [a,*], Kiyoaki Shirai [a], Julien Velcin [b]

[a] School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
[b] University of Lyon (ERIC, Lyon 2), 5 Avenue Pierre Mendes-France, 69676 Bron Cedex, France

ARTICLE INFO

Keywords:
Sentiment analysis
Opinion mining
Classification
Prediction
Stock
Social media
Message board

ABSTRACT

The goal of this research is to build a model to predict stock price movement using the sentiment from social media. Unlike previous approaches where the overall moods or sentiments are considered, the sentiments of the specific topics of the company are incorporated into the stock prediction model. Topics and related sentiments are automatically extracted from the texts in a message board by using our proposed method as well as existing topic models. In addition, this paper shows an evaluation of the effectiveness of the sentiment analysis in the stock prediction task via a large scale experiment. Comparing the accuracy average over 18 stocks in one year transaction, our method achieved 2.07% better performance than the model using historical prices only. Furthermore, when comparing the methods only for the stocks that are difficult to predict, our method achieved 9.83% better accuracy than historical price method, and 3.03% better than human sentiment method.

1. Introduction

Stock price forecasting is very important in the planning of business activity. However, building an accurate stock prediction model is still a challenging problem. In addition to historical prices, the current stock market is affected by the mood of society. The overall social mood with respect to a given company might be one of the important variables which affect the stock price of that company. Nowadays, the emergence of online social networks makes available large amounts of mood data. Therefore, incorporating information from social media with the historical prices can improve the predictive ability of models.

The goal of our research is to develop a model to predict the stock price movement (whether the price will be up or down) using information from social media (Message Board). In our proposed method, a model that predicts the stock value at t using features derived from information at $t-1$ and $t-2$, where t stands for a transaction date, will be trained by supervised machine learning. Apart from the mood information, the stock prices are affected by many factors such as microeconomic and macroeconomic factors. However, this research only focuses on how the mood information from social media can be used to predict the stock price. We will mainly aim at extracting the mood information by sentiment analysis on social media data. Then, these sentiments will be integrated into a model to predict stocks. To achieve this goal, discovering the topics and sentiments in a large amount of social media is very important to get the opinions of investors. However, sentiment analysis on social media is difficult. The text is usually short, contains many misspellings, uncommon grammar constructions and so on. In addition, the literature shows conflicting results in sentiment analysis for stock market prediction. Some researchers report that sentiments from social media have no predictive capabilities (Antweiler & Frank, 2004; Tumarkin & Whitelaw, 2001), while other researchers have reported either weak or strong predictive capabilities (Bollen, Mao, & Zeng, 2011). Therefore, how to use opinions in social media for stock price predictions is still an open problem.

One contribution of this paper is that we propose a novel feature 'topic-sentiment' to improve the performance of stock market prediction. It is important to recognize what topics are discussed in social media and how people feel about these topics. The 'topic-sentiment' feature, which represents the sentiments of the specific topics of the company (product, service, dividend and so on), are used for prediction of stock price movement. This feature is obtained in two ways: by using the existing topic model called the joint sentiment/topic model (JST) and by our own proposed method. The extracted topics and sentiments in the former method are hidden (latent), whereas not hidden in the latter. To the best of our knowledge, this is the first research trying to extract topics and sentiments simultaneously and utilize them for stock market prediction. Another contribution is a large scale evaluation. The effectiveness of the sentiments in social media in stock market prediction is still uncertain because a

* Corresponding author. Tel.: +81 80 2956 5927.
E-mail addresses: nhthien8x@gmail.com, nhthien@jaist.ac.jp (T.H. Nguyen), kshirai@jaist.ac.jp (K. Shirai), julien.velcin@univ-lyon2.fr (J. Velcin).

http://dx.doi.org/10.1016/j.eswa.2015.07.052
0957-4174/

relatively small data was used for evaluation in the previous work. This paper investigates whether the sentiments in the social media are really useful on the test data containing many stocks and transaction dates.

The rest of the paper is organized as follows. Section 2 introduces some previous approaches on sentiment analysis for stock prediction. Section 3 describes our dataset. Section 4 describes our proposed method. We also propose a novel feature for stock prediction based on the topics and the sentiments associated with them. Section 5 assesses the results of the experiments. Finally, Section 6 concludes our contribution.

2. Related work

Stock market prediction is one of the most attracted topics in academic as well as real life business. Many researches have tried to address the question whether the stock market can be predicted. Some of the researches were based on the random walk theory and the Efficient Market Hypothesis (EMH). According to the EMH (Fama, 1991; Fama, Fisher, Jensen, & Roll, 1969), the current stock market fully reflects all available information. Hence, price changes are merely due to new information or news. Because news in nature happens randomly and is unknowable in the present, stock prices should follow a random walk pattern and the best bet for the next price is current price. Therefore, they are not predictable with more than about 50% accuracy (Walczak, 2001). On the other hand, various researches specify that the stock market prices do not follow a random walk, and can be predicted at some degree (Bollen et al., 2011; Qian & Rasheed, 2007; Vu, Chang, Ha, & Collier, 2012). Degrees of directional accuracy at 56% hit rate in the predictions are often reported as satisfying results for stock predictions (Schumaker & Chen, 2009b; Si, Mukherjee, Liu, Li, Li, & Deng, 2013; Tsibouris & Zeidenberg, 1995).

Besides the efficient market hypothesis and the random walk theories, there are two distinct trading philosophies for stock market prediction: fundamental analysis and technical analysis. The fundamental analysis studies the company's financial conditions, operations, macroeconomic indicators to predict stock price. On the other hand, the technical analysis depends on historical and time-series prices. Price moves in trends, and history tends to repeat itself. Some researches have tried to use only historical prices to predict the stock price (Cervelló-Royo, Guijarro, & Michniuk, 2015; Patel, Shah, Thakkar, & Kotecha, 2015a, 2015b; Ticknor, 2013; Zuo & Kita, 2012a, 2012b). To discover the pattern in the data, they used Bayesian network (Zuo & Kita, 2012a, 2012b), time-series method such as Auto Regressive model, Moving Average model (Patel et al., 2015a, 2015b), Auto Regressive Moving Average model (Zuo & Kita, 2012a) and so on.

While these previous methods did not consider the sentiments on the social media, in this paper our work aims at incorporating them to improve the performance of the stock market prediction.

Most of the research tried to predict only one stock (Bollen et al., 2011; Qian & Rasheed, 2007; Si et al., 2013) and the number of instances (transaction dates) in a test set is very low such as 14 or 15 instances (Bollen et al., 2011; Vu et al., 2012). With only a few instances in the test set, the conclusion might be insufficient. To the best of our knowledge, there is no research showing a good prediction result on a data consisting of many stocks in a long time period. Our research tried to solve this issue by predicting 18 stocks over a period of one year.

2.1. Use of opinions from text for stock market prediction

Sentiment analysis has been found to play a significant role in many applications such as product reviews and restaurant reviews (Liu & Zhang, 2012; Pang & Lee, 2008). There are some researches trying to apply sentiment analysis on an information source to improve the stock prediction model (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014). There are two main sources from which authors have incorporated information aggregated from textual content into financial models. In the past, the main source was the news (Schumaker & Chen, 2009a, 2009b), and in recent years, social media sources. Then, these sentiments are integrated into prediction models. A simple approach is combining the textual content with the historical prices through the linear regression model.

Most of the previous work primarily used the bag-of-words as text representation that are incorporated into the prediction model. Schumaker and Chen (2009b) tried to use different textual representations such as bag-of-words, noun phrases and named entities for financial news. Then this information was integrated with linear regression and support vector machine regression as predictive models. They applied their models to estimate a discrete stock price 20 min. after a news article was released. The results show 0.04261 mean square error, 57.1% directional accuracy, and 2.06% return in a simulated trading engine. However, the textual representations are just the words or named entity tags, not exploiting so much about the mood information.

Antweiler and Frank (2004) used naive Bayes to classify the messages from message boards into three classes: buy, hold and sell. The number of relevant messages in these three classes was aggregated into a single measure of bullishness. They investigated three aggregation functions as a number of alternatives to bullishness. They were integrated into the regression model. However, they concluded that their model does not successfully predict stock returns.

Zhang, Fuehres, and Gloor (2011) measured collective hope and fear on each day and analyzed the correlation between these indices and the stock market indicators. They used the mood words to tag each tweet as fear, worry, hope and so on. They concluded that the emotional tweet percentage significantly negatively correlated with Down Jones, NASDAQ and S&P 500, but had significant positive correlation to VIX. However, they did not use their model to predict the stock price values.

Two mood tracking tools, OpinionFinder and Google Profile of Mood States, were used to analyze the text content of daily Twitter (Bollen et al., 2011). The former measures positive and negative mood. The latter measures mood in terms of six dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). They used the Self Organizing Fuzzy Neural Network model to predict DJIA values. The results show 86.7% direction accuracy (up or down), Mean Absolute Percentage Error 1.79%. However, their test period is very short (from December 1 to December 19, 2008). Even though, they achieved high accuracy, there are only 15 transaction dates in their test set. With such a short period, it might not be sufficient to conclude the effectiveness of their method.

Xie, Passonneau, Wu, and Creamer (2013) proposed a novel tree representation based on semantic frame parsers. They indicated that this representation performed significantly better than bag-of-words. By using stock prices from Yahoo Finance, they annotated all the news with labels in a transaction date as going up or down categories. However, the weakness of this assumption is that all the news in one day will have the same category. In addition, this becomes a document classification problem, not stock prediction.

Rechenthin, Street, and Srinivasan (2013) incorporated Yahoo Finance Message Board into the stock movement prediction. They tried to use various classification models to predict stock. They used the explicit sentiments and predicted sentiments obtained by a classification model with the bag-of-words and meta-features.

A keyword-based algorithm was proposed to identify the sentiment of tweets as positive, neutral and negative for stock prediction (Vu et al., 2012). Their model achieved around 75% accuracy. However, their test period is very short, from 8th to 26th in September, 2012 which contains only 14 transaction dates.

Si et al. (2013) developed a non-parametric topic model for Twitter messages to predict the stock market. They proposed a continuous Dirichlet Process Mixture (cDPM) model to learn the daily topic set. Then, a sentiment time series was built based on these topics. The advantage of this method is that the model estimates the number of topics inherent in the data itself. However, the time period of their dataset is quite short, only three months.

A series of the previous work discussed in this subsection tried to extract the overall opinion or sentiment of the document. However, the opinions are often expressed toward the topics or aspects. For the prediction of stock prices, it is important to know on which topics of the company people have positive or negative opinions. In our proposed method, the sentiments of the topics or aspects are extracted, then they are incorporated into the stock prediction models.

Next subsection will discuss some related work for the identification of aspect-oriented sentiments.

2.2. Aspect based sentiment analysis

There are some researches trying to extract both topic and sentiment for other domains such as online product review, restaurant review and movie review dataset (Dermouche, Kouas, Velcin, & Loudcher, 2015). Jo and Oh (2011) proposed ASUM model for extracting both aspect and sentiment for online product review dataset. The model assumes that all words within a sentence are generated from one topic.

The joint sentiment/topic model (JST) was proposed to detect sentiment and topic simultaneously for movie review dataset (Lin & He, 2009). This model assumes that each word is generated from a joint topic and sentiment distribution. Because this model can extract topic and sentiment simultaneously, we will use it to extract topic-sentiment features.

Lakkaraju, Bhattacharyya, Bhattacharya, and Merugu (2011) proposed the FACTS, CFACTS, FACTS-R, and CFACTS-R model to perform sentiment analysis on a product review data. These models assume a word fallen into three categories: facet word, sentiment word or other category (background word, stop word, function word, etc.). Based on its category, a word is generated from corresponding facet, sentiment or background distribution. In addition, they introduced a *window*, which is a contiguous sequence of words. All facet words within a window are assumed to be derived from the same facet topic, and all sentiment words from the same sentiment topic.

Zhao, Jiang, Yan, and Li (2010) proposed the MaxEnt-LDA hybrid model to jointly discover both aspects and aspect-specific opinion words on a restaurant review dataset. Besides the general opinion words, they considered the aspect-specific opinion words. Therefore, a word is fallen into five categories: background, specific aspect, general aspect, specific opinion and general opinion. Based on these categories, words are generated from corresponding distributions.

The above methods tried to extract the hidden (latent) topic-sentiment associations. In our proposed method, both hidden and non-hidden topic-sentiment are considered by using JST topic model and proposed algorithms that will be discussed in Sections 4.5 and 4.6, respectively.

3. Dataset

We used two datasets for our stock prediction model. The first one is the historical price dataset, and the second one is the mood information dataset.

3.1. Historical prices

Historical prices are extracted from Yahoo Finance for the 18 stocks. The list of stock quotes and company names is shown in the Table 1. For each transaction date, there are open, high, low, close and adjusted close prices. The adjusted close prices are the close prices which are adjusted for dividends and splits. The adjusted close price is often used for stock market prediction as in other researches (Rechenthin et al., 2013). Therefore, we chose it as the stock price value for each transaction date.

Table 1
Quotes and company names.

Stocks	Company names
AAPL	Apple Inc.
AMZN	Amazon.com Inc.
BA	The Boeing Company
BAC	Bank of America Corporation
CSCO	Cisco Systems Inc.
DELL	Dell Inc.
EBAY	eBay Inc.
ETFC	E Trade Financial Corporation
GOOG	Google Inc.
IBM	International Business Machines Corporation
INTC	Intel Corporation
KO	The Coca-Cola Company
MSFT	Microsoft Corporation
NVDA	NVIDIA Corporation
ORCL	Oracle Corporation
T	AT&T Inc.
XOM	Exxon Mobil Corporation
YHOO	Yahoo! Inc.

3.2. Message board dataset

To get the mood information of the stocks, we collected the 18 message boards of the 18 stocks from Yahoo Finance Message Board for a period of one year (from July 23, 2012 to July 19, 2013)[1]. On the message boards, users usually discuss company news, prediction about stock going up or down, facts, comments (usually negative) about specific company executives or company events. In 15.6% messages in this dataset, when users posted messages on these message boards, they annotated each message as one of the following sentiment tags: Strong Buy, Buy, Hold, Sell and Strong Sell. There are two kinds of messages. The first one is the messages created by starting a new topic. The other is reply messages to existing messages. Most of users' posts are reply messages. They form a complicated communication network. In our research, however, we treated all messages independent from each other.

Fig. 1 shows an example message from AAPL Message Board. In this message, on July 6, 2012 a username "keepshorting" posted the message "Looks like the competition is heating up. $199 tablet, what is next? $999 laptops and then $499 laptops? the margins are impossible to keep up. impossible folks." to reply to another message of another user. In addition, this user selected the sentiment for this stock as "Strong Sell".

The stock market is not opened at the weekend and holiday. To assign the messages to the transaction dates, the messages which were posted from 4 pm of the previous transaction date to 4 pm of the current transaction date will belong to the current transaction. We choose 4 pm because that is the time of closing transaction. There are 249 transaction dates from the one year period of our dataset. Table 2 summarizes the statistics of our dataset for each transaction date about the min, median, mean, max of the number of messages and the mean of the number of the existing sentiments annotated by users.

Some previous works used Twitter as the mood information source for sentiment analysis related to a particular stock. There are

[1] The AAPL message board has the highest number of messages. Because of the limitation on the number of web pages, we can only collect for a period of seven months for this stock.

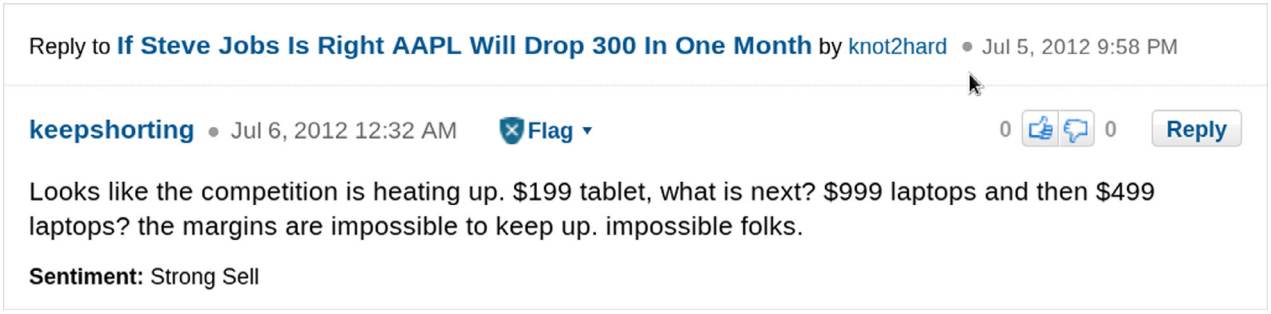

Fig. 1. A message from AAPL message board.

Table 2
Statistics of our dataset for each transaction date.

Stocks	The number of messages				Mean of the number of human sentiments
	Min	Median	Mean	Max	
AAPL	0	1093	1678	11220	350
AMZN	24	154	192	1963	28
BA	46	173	203	1053	16
BAC	94	282	343	1366	49
CSCO	69	247	274	972	10
DELL	0	18	42	587	10
EBAY	1	17	29	267	3
ETFC	2	42	56	315	12
GOOG	10	69	93	1305	16
IBM	3	14	20	195	3
INTC	37	177	200	958	29
KO	0	6	8	89	2
MSFT	27	139	172	815	53
NVDA	10	65	80	410	11
ORCL	5	67	79	372	6
T	10	52	59	251	8
XOM	10	37	44	202	4
YHOO	22	121	141	860	27

some reasons why in our research Twitter is not chosen as a mood source. The first one is the information in Twitter seems to be messier than that in the Message Board. In the Twitter, users discuss about many things. Even though tweets can be filtered by some rules such as using hashtag (#AAPL, $AAPL and so on) to find relevant tweets, the lack of consistency among posters in hashtag use might be problematic. Furthermore, the high level of noise makes finding a post related to a specific stock difficult. The second reason is the way to collect tweets. There are two ways to collect tweets from the Twitter. The first one is from the Twittter Searching API. This only allows searching tweets from one week in the past for free. The other way is using Twitter Streaming API. It allows collecting the real time tweets rather than search from the history. However, to collect tweets in one year period, it takes one year. Those make difficult to gather a large amount of data from the Twitter. Finally, there are no explicit sentiments annotated by posters in the Twitter. There is no way to compare between human sentiment and automatic sentiment extraction.

However, as in other mood information sources, the messages on the Message Board are also messy. The text is usually short, contains many misspellings, uncommon grammar constructions and so on. Moreover, the false and unrelated information also exists.

4. Methods for stock movement prediction

The Support Vector Machine (SVM) has long been recognized as being able to efficiently handle high dimensional data and has been shown to perform well on classification (Joachims, 1998; Nguyen & Shirai, 2013). Therefore, we chose the SVM with the linear kernel as the prediction model. To assess the effectiveness of sentiment analysis on the message boards, six sets of features are designed. The first one used only the historical prices. The other methods incorporated the mood information into the prediction model. All the feature values were scaled into $[-1, 1]$. Table 3 summarizes our features used in the prediction model to predict the price movement at the transaction date t. The details of each feature will be explained in the next subsections.

Table 3
Features of the prediction model.

Method	Features
Price only	$price_{t-1}$, $price_{t-2}$
Human sentiment	$price_{t-1}$, $price_{t-2}$, $Hsent_{i,t}$, $Hsent_{i,t-1}$
Sentiment classification	$price_{t-1}$, $price_{t-2}$, $Csent_{i,t}$, $Csent_{i,t-1}$
LDA-based method	$price_{t-1}$, $price_{t-2}$, $lda_{i,t}$, $lda_{i,t-1}$
JST-based method	$price_{t-1}$, $price_{t-2}$, $jst_{i,j,t}$, $jst_{i,j,t-1}$
Aspect-based sentiment	$price_{t-1}$, $price_{t-2}$, $Asent_{i,t}$, $Asent_{i,t-1}$, $l_{i,t}$, $l_{i,t-1}$

4.1. Price only

In this method, only historical prices are used to predict the stock movement. The purpose of this method is to investigate whether there are patterns in the history of the stock or not. In addition, this model was used as a baseline to evaluate whether integration of the sentiments is effective by comparing with other sentiment models. Features used for the training of SVM are $price_{t-1}$ and $price_{t-2}$ which are the price movements (up, down) at the transaction dates $t-1$ and $t-2$, respectively.

4.2. Human sentiment

In addition to historical prices, this model integrated the sentiments annotated by human into the prediction model. As discussed in Section 3.2, in 15.6% of the MessageBoard dataset, the users explicitly select a sentiment label with their posts. These sentiment labels are "strong buy", "buy", "hold", "sell" and "strong sell". Instead of using all the messages, we tried to use only the messages with annotated sentiments by the users, and discard the other messages. From these messages, we used only the explicit sentiment and remove other information such as message content. The purpose of this method is that how mood annotated by human can be used to predict the stock. Because the sentiments are annotated by human, this feature is one of the strongest features for stock prediction.

For each transaction date t, the percentage of each class (Strong Buy, Buy, Hold, Sell, and Strong Sell) was calculated. The percentage of a class is the number of messages having sentiments as that class label divided by the number of messages in the current transaction date t. Then, we integrated them into the prediction model.

Features used for training SVM are $price_{t-1}$, $price_{t-2}$, $Hsent_{i,t}$ and $Hsent_{i,t-1}$. $Hsent_{i,t}$ and $Hsent_{i,t-1}$ are the percentages of the number of messages belonging to the sentiment class i ($i \in$ {Strong Buy,

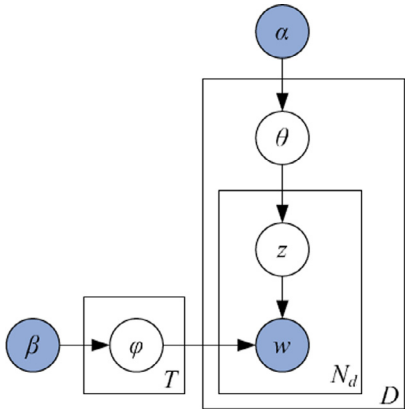

Fig. 2. Graphical model representation of LDA.

Buy, Hold, Sell, and Strong Sell}) at the transaction dates t and $t-1$, respectively.

4.3. Sentiment classification

To utilize the remaining 84.4% of the messages without the explicit sentiments, we tried to build a model to extract the sentiments for those messages. A classification model was trained from the messages with annotated sentiments on the training dataset. Then it was used to classify the remaining messages into five classes (Strong Buy, Buy, Hold, Sell, and Strong Sell).

We removed the stop words from the messages. Then, all the words are lemmatized by the Stanford CoreNLP (Manning, Surdeanu, Bauer, Finkel, Bethard, & McClosky, 2014). The feature representation is the bag-of-words from the title and content of the message. The feature weighting is TF-IDF. We chose SVM with the linear kernel as the classification model.

As in the human sentiment feature, we also calculated the percentage of the number of messages of each class for each transaction date. Features used for the training of SVM are $price_{t-1}$, $price_{t-2}$, $Csent_{i,t}$ and $Csent_{i,t-1}$. $Csent_{i,t}$ and $Csent_{i,t-1}$ are similar to $Hsent_{i,t}$ and $Hsent_{i,t-1}$, but both messages with human annotated sentiment and automatically classified sentiments are used to calculate the percentages of the number of messages belonging to the sentiment class i.

4.4. LDA-based method

In this model, we consider each message as a mixture of hidden topics. Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Therefore, we choose the LDA as a simple topic model to discover these hidden topics[2]. Fig. 2 shows the graphical model representation of LDA. Notations in Fig. 2 are shown in Table 4.

We removed the stop words from messages. Then, all the words are lemmatized by the Stanford CoreNLP. We train the LDA on the training set, and infer the topics for unseen messages on the test set. Topics are inferred by the Gibbs Sampling with 1000 iterations. We chose 50 as the number of topics. After that, the probability of each topic for each message is calculated. Next, for each transaction date t, the probability of each topic is defined as the average of the probabilities of that topic in the messages belonging to that transaction date. Then we integrated these probabilities into the prediction model.

[2] We used the LDA implementation from the Mallet library.

Table 4
Notations in LDA.

Notation	Definition
α, β	Hyperparameters
φ	The distribution over words
T	The number of topics
θ	The message specific topic distribution
z	A topic
w	A word in the message d
N_d	The number of words in the message d
D	The number of messages

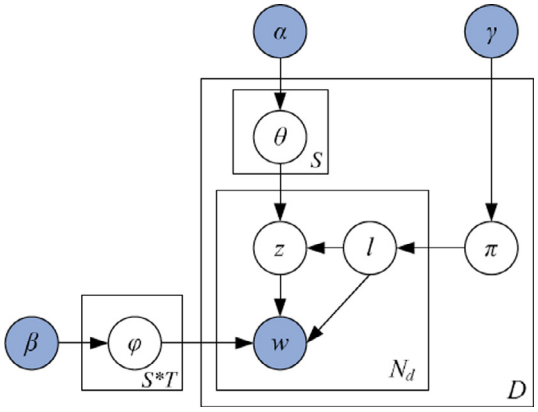

Fig. 3. Graphical model representation of JST.

Table 5
Notations in JST.

Notation	Definition
α, β, γ	Hyperparameters
φ	The distribution over words
T	The number of topics
S	The number of sentiments
θ	The message and sentiment specific topic distribution
z	A topic
w	A word in the message d
l	A sentiment label
π	The message specific sentiment distribution
N_d	The number of words in the message d
D	The number of messages

Features used for training SVM are $price_{t-1}$, $price_{t-2}$, $lda_{i,t}$ and $lda_{i,t-1}$. $lda_{i,t}$ and $lda_{i,t-1}$ are the probabilities of the topic i ($i \in \{1, \ldots, 50\}$) for the transaction dates t and $t-1$.

4.5. JST-based method

The opinion is often expressed on a topic or aspect. When people post the message on the social media to express their opinion for a given stock, they tend to talk their opinions for a certain topic such as profit and dividend. Based on pairs of topic-sentiment, they would think that the future price of that stock goes up or down. From that intuition, we propose a new feature topic-sentiment for the stock prediction model. To extract pairs of topic-sentiment, we tried to use two kinds of models. The first one is a latent topic based model, the JST model (Lin & He, 2009). The second one is Aspect-based Sentiment model which will be discussed in the next Section 4.6.

We consider each message as a mixture of hidden topics and sentiments. The JST model was used to extract topics and sentiments simultaneously. Fig. 3 shows the graphical model representation of JST. Notations in Fig. 3 are shown in Table 5. In LDA model, there is only one document specific topic distribution for each document. In contrast, each document in JST is associated with S sentiment labels. Each

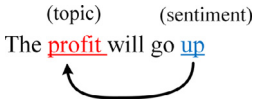

Fig. 4. An example sentence with topic and its sentiment.

of sentiment labels is associated with a document specific topic distribution with the same number of topics. A word in the document is drawn from distribution over the words defined by the topic and sentiment label.

After removal of stop words and lemmatization, the JST model is trained from the training set, and topics on the test set are inferred by the Gibbs Sampling with 1000 iterations. We chose 50 as the number of topics and 3 as the number of sentiments. Next, the joint probability of each pair of topic and sentiment is calculated for each message. After that, for each transaction date t, the joint probability of each topic-sentiment pair is defined as the average of the joint probabilities of that in the messages belonging to that transaction date. Then we integrated these probabilities into the prediction model.

Features used for training SVM are $price_{t-1}$, $price_{t-2}$, $jst_{i,j,t}$ and $jst_{i,j,t-1}$. $jst_{i,j,t}$ and $jst_{i,j,t-1}$ are the joint probabilities of the sentiment i ($i \in \{1, 2, 3\}$) and topic j ($j \in \{1, \ldots, 50\}$) for the transaction dates t and $t-1$.

4.6. Aspect-based sentiment

Instead of considering the mixtures of hidden topics and sentiments as in the previous model, in this model the mixtures are not hidden. Each message is represented as a list of topics and their corresponding sentiment values. In our proposed method, the topic is the consecutive nouns in the sentence. For example, the message "The profit will go up." contains the topic "profit" and a positive sentiment "up" for that topic as in Fig. 4.

We propose a new model to calculate the sentiment values of the topics in a sentence. For each message, the sentences are split. Then, we used the Stanford CoreNLP for POS tagging and lemmatization of each word in each sentence. First, we extracted the topics in the training dataset by using the algorithm shown in Fig. 5. We extract the consecutive nouns as the topics in the sentence. To eliminate rare topics, topics occurring less than 10 times are removed from the list of the topics. Next, based on the topic list, we extracted their sentiment values in each sentence by using the algorithm shown in Fig. 6. For each sentence, opinion words are identified based on the list of opinions from SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010). SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns each word three sentiment scores: positivity, objectivity and negativity. We combined scores of positivity and negativity into a single opinion value. The closer between the topic phrase and the opinion word, the higher affection of that opinion on the topic phrase. Therefore, the sentiment value of a topic phrase in a sentence is the summation of overall opinion values divided by their distance to that topic.

For each message, the sentiment value of each topic is defined as the average of the sentiment scores of that topic in the sentences. Finally, for each transaction date t, the sentiment value for each topic is defined as the average of the sentiment

Input: Training dataset
Output: List of topics of this dataset
1 Extract consecutive nouns in each sentence as a topic ;
2 Remove topics that appear less than 10 times in the training dataset ;

Fig. 5. Algorithm for extracting topics from dataset.

Input: A sentence
Output: List of pairs $(topic, sentimentValue)$ for this sentence
1 Extract topics in the sentence (Based on the list of topics extracted from the algorithm shown in Figure 5) ;
2 Extract opinion words in the sentence by using SentiWordNet ;
3 **for** *each topic* t_i *in the sentence* **do**
4 **for** *each opinion* o_j *in the sentence* **do**
5 Calculate $distance(t_i, o_j)$ = position distance between topic t_i and opinion word o_j ;
6 Get pos_score, neg_score of opinion o_j from SentiWordNet ;
7 Calculate $opinionValues(o_j) = \frac{pos\_score - neg\_score}{pos\_score + neg\_score}$;
8 $sentimentValue_{t_i} = sentimentValue_{t_i} + \frac{opinionValues(o_j)}{distance(t_i, o_j)}$;
9 Add $(t_i, sentimentValue_{t_i})$ to the list of pairs $(topic, sentimentValue)$
10 **end**
11 **end**

Fig. 6. Algorithm for extracting topics and their sentiment values.

Table 6
Results of accuracies of 18 stocks.

Stocks	Baseline models				Our models	
	Price only	Human sentiment	Sentiment classification	LDA-based method	JST-based method	Aspect-based sentiment
AAPL	0.3951	0.5679	0.4938	0.5802	0.5802	0.5432
AMZN	0.4605	0.4868	0.4605	0.5132	0.5921	0.7105
BA	0.6316	0.6053	0.5132	0.5526	0.6316	0.5921
BAC	0.5658	0.5921	0.5658	0.5526	0.5658	0.4474
CSCO	0.5526	0.4474	0.5263	0.4737	0.5132	0.4605
DELL	0.5395	0.5921	0.4737	0.5132	0.4342	0.6447
EBAY	0.5921	0.4605	0.4605	0.5658	0.4079	0.5789
ETFC	0.5789	0.5921	0.5789	0.4868	0.4342	0.5526
GOOG	0.5000	0.5658	0.5789	0.5658	0.5395	0.5263
IBM	0.4868	0.4737	0.4868	0.5395	0.4474	0.5526
INTC	0.4474	0.4605	0.4342	0.5000	0.4868	0.5263
KO	0.4079	0.4868	0.5132	0.5658	0.5132	0.4474
MSFT	0.5789	0.6579	0.5921	0.5526	0.5526	0.5263
NVDA	0.6053	0.5789	0.6184	0.3947	0.5000	0.5395
ORCL	0.4868	0.5263	0.5263	0.5921	0.5000	0.5395
T	0.5526	0.4737	0.4868	0.5000	0.5658	0.5132
XOM	0.4868	0.6447	0.4868	0.4342	0.5658	0.5395
YHOO	0.5526	0.5526	0.5395	0.5263	0.4474	0.5526
AVERAGE	0.5234	0.5425	0.5187	0.5227	0.5154	**0.5441**

values of that topic in the messages belonging to that transaction date.

In addition to the sentiment values of the topics, the importance of the topics for each transaction date were also considered. Intuitively, some topics have more impact on the prediction than others. If a topic was discussed in many messages, it might be an important topic in the given transaction date. The importance of a topic i in a transaction date t was calculated as in Eq. (1). It is defined as the fraction between the number of messages containing the current topic i in the transaction date t and the number of messages in that transaction date.

$$I_{i;t} = \frac{N_{i;t}}{N_t} \tag{1}$$

where:

$I_{i;t}$: the importance of topic i in the transaction date t.
$N_{i;t}$: the number of messages containing the topic i in the transaction date t.
N_t: the number of messages in the transaction date t.

The sentiment scores of the topics at the transaction date level and their important were used in the prediction model. Features used for training SVM are $price_{t-1}$, $price_{t-2}$, $Asent_{i;t}$, $Asent_{i,t-1}$, $I_{i;t}$ and $I_{i;t-1}$. $Asent_{i;t}$ and $Asent_{i,t-1}$ are the sentiment values of the topic i at the transaction dates t and $t-1$. While, $I_{i;t}$, $I_{i;t-1}$ are the importance of the topic i at the transaction dates t and $t-1$.

5. Evaluation

5.1. Experiment setup

We divided the time series into two parts: the period from July 23, 2012 to March 28, 2013 for training containing 171 transaction dates, and April 01, 2013 to July 19, 2013 for testing containing 78 transaction dates[3]. We assigned each transaction date a label (up, down) by comparing its price with the previous transaction date's price. The performance is evaluated using the Accuracy metric. Accuracy is the proportion of true results in the test set.

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \tag{2}$$

where:

tp: the number of samples correctly categorized for positive samples.
tn: the number of samples correctly rejected for the negative samples.
fp: the number of samples incorrectly categorized for the positive samples.
fn: the number of samples incorrectly rejected for the negative samples.

5.2. Experiment results

The results of accuracy measure are shown in Table 6. In addition to the result of each stock, we also calculated the average of 18 stocks for each model for easy comparison. Using Aspect-based sentiment feature achieved the best result with 54.41% average accuracy for 18 stocks. As discussed in Section 2, degrees of accuracy of 56% hit rate are often reported as satisfying results for stock prediction. In addition, the number of instances (transaction dates) in test set of most of other researches is small, and the number of stock is usually only one. In contrast, the advantage of this work is that we used the training and test data on a long period (one year) containing many instances, and for many stocks (18 stocks). For some stocks, the accuracies are quite high, such as 71.05% for AMZN stock, 64.47% for DELL stock and so on.

To assess the effectiveness of integrating mood information, we compare our Aspect-based sentiment method with the Price only method. The results show that the model using mood information outperforms 2.07% on the average accuracy than the model without mood. Furthermore, comparing the Human sentiment with Price only method, it indicated that the prediction accuracy is improved 1.91% by using the sentiments annotated by human. Therefore, we can conclude that integration of the sentiments of both overall documents and specific topics from social media could help to improve the stock market prediction.

To assess the effectiveness of automatic sentiment analysis and human sentiment, we compare our Aspect-based sentiment method with the Human sentiment method. The results show that our automatically extracted sentiment is slightly higher than using the sentiment annotated by human. Therefore, our method is comparable to the human sentiment method. Note that the advantage of our method

[3] For AAPL: the period July 06–October 01, 2012 for training containing 61 transaction dates, and November 12, 2012–March 13, 2013 for test containing 83 transaction dates.

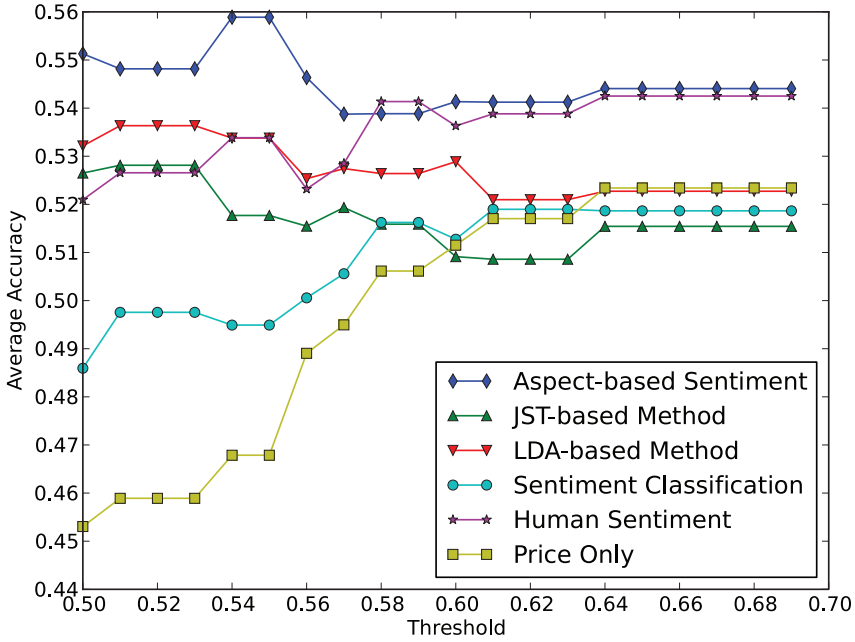

Fig. 7. Comparison of the models for different threshold α.

is that it can be applicable for other social media without human annotated sentiment such as the Twitter.

The Aspect-based sentiment method outperformed over 2.54%, 2.14% and 2.87% on average accuracy compared to Sentiment classification, LDA-based method and JST-based method, respectively. The LDA-based and JST-based method seem to be not successful in this experiment. The limitation of these methods is that we have to specify the number of hidden topics in LDA and the number of hidden topics and sentiments in JST. In our experiment, we specify 50 topics for all of the stock. This assumption would be not appropriate in general. For the individual stock, the number of discussed topics depends on the content of the messages. Therefore, the appropriate number of hidden topics may be varied for different stocks. However, there is no way to determine the number of topics in the model of LDA and JST. One of the solutions is a grid search trying different number of topics and finding the best value. However, since the running time of the Gibbs Sampling depends on the size of the dataset, it takes very long time to run it repeatedly on a big dataset of 18 stocks for a long period. Therefore, a grid search cannot be tested in our experiment.

To assess the effectiveness of topic-sentiment feature, we compare our Aspect-based sentiment method with Sentiment classification method (using only sentiment information) and LDA-based method (using only topic information). The results indicate that using topic-sentiment feature is better than using only sentiments by 2.54% accuracy, and using only topics by 2.14% accuracy. Therefore, understanding on which topics the sentiments are expressed is useful for stock market prediction. In other words, the topic-sentiment feature is better than only topic or sentiment feature.

Although the sentiment information is effective for the stock prediction on average, in the comparison on the individual stocks, the model with sentiment analysis is worse than the price only model for several stocks. There are many possible reasons for it. As discussed in Section 2, the stock market is influenced by many factors. Some proposed that they are not predictable with more than 50% accuracy. One reason is that the sentiment might not be a factor which causes the stock price moving. Another reason is that even though sentiment might be one of the factors which affect price moving, the extracted sentiments from the Message Boards do not reflect the price because of the messy, fault comment or fault prediction of human when they post the messages.

A simple assumption about the effectiveness of sentiment feature is that the sentiment analysis may not provide any additional information if the stock movement can be predicted well by the historical price only. If the accuracy of the price only model is high, there are trends and historical repetition in the stock. In such cases, only historical prices might be enough to predict, and integration of the sentiment may not improve the accuracy much. On the other hand, if the accuracy of the price only model is low, the stock seems to have no pattern in its history. For such stocks, the use of sentiment may be effective for the prediction.

To investigate the above assumption, we compare the models from another point of view. First, we define a threshold α. If the accuracy of the stock in the Price only method ($A_{PriceOnly}$) is higher than α, this stock is discarded from the evaluation. In other words, we compared the average accuracy for the stocks where $A_{PriceOnly} < \alpha$. Fig. 7 shows the average accuracies against various thresholds. It is found that the difference between the models with and without sentiment information becomes greater when α is set smaller. At the threshold 50%, using our Aspect-based sentiment model improved the accuracy over 9.83% compared to Price only, over 3.03% compared to Human sentiment method. In addition, in most of the thresholds, our method achieved the best accuracy compared with other methods.

6. Conclusion & future work

Stock price prediction is a challenging task because the stock prices are affected by many factors. This paper presents the novel method to integrate the sentiments in social media for the prediction of stock price movement. The contribution of this study can be summarized as follows. First, while the overall sentiments in the documents are considered in the previous research, this research proposed a method using the sentiment of the topic for stock market prediction. Second, we proposed two methods to capture these topic-sentiment associations. One is JST-based method that relies on the existing topic model, the other is Aspect-based sentiment method where the topics and sentiments are identified by the proposed method. Finally, this is the first research to show the effectiveness of incorporation of the sentiment analysis by investigation on a large scale test data. From a practical point of view, although the average accuracy is only 54.41%, the proposed method can predict the stock

price movement with more than 60% accuracy for a few stocks, and performs much better than other methods for the stocks that are difficult to predict with only past prices.

A limitation of this research is that we specified the number of topics and sentiment beforehand for the LDA and JST-based method. To overcome this weakness, a non-parametric topic model that can infer the number of topics and sentiments automatically is useful to extract the topic and sentiment simultaneously for the stock prediction. This will be done in our future work.

The current model only predicts if the stock price is up or down. However, people may want to forecast drastic movement of the stock market. In that sense, the proposed model is insufficient. However, our model can be extended to predict the degree of the change by setting more fine grained classes such as 'great up', 'little up', 'little down', 'great down' and so on.

One of the weaknesses of our method is that only the historical prices and sentiments derived from social media are considered. In future, we will try to find and integrate more factors which can affect the stock prices to develop a more accurate stock prediction model. For example, co-variance between stocks, macroeconomic indicators and the financial conditions of the company, which can be guessed from the income statement, balance sheet and cash flow, are important factors to be considered in the stock prediction model.

References

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

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the seventh conference on international language resources and evaluation: vol. 10* (pp. 2200–2204).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8.

Cervelló-Royo, R., Guijarro, F., & Michniuk, K. (2015). Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA index with intraday data. *Expert Systems with Applications, 42*(14), 5963–5975.

Dermouche, M., Kouas, L., Velcin, J., & Loudcher, S. (2015). A joint model for topic-sentiment modeling from text. In *ACM/SIGAPP symposium on applied computing (sac)* (pp. 819–824).

Fama, E. F. (1991). Efficient capital markets: II. *The Journal of Finance, 46*(5), 1575–1617.

Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The adjustment of stock prices to new information. *International Economic Review, 10*(1), 1–21.

Jo, Y., & Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on web search and data mining* (pp. 815–824). ACM.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. Springer.

Lakkaraju, H., Bhattacharyya, C., Bhattacharya, I., & Merugu, S. (2011). Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In *Proceedings of the eleventh SIAM international conference on data mining* (pp. 498–509). SIAM /Omnipress.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on information and knowledge management* (pp. 375–384). ACM.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415–463). Springer.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55–60).

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: a systematic review. *Expert Systems with Applications, 41*(16), 7653–7670.

Nguyen, T. H., & Shirai, K. (2013). Text classification of technical papers based on text segmentation. In *Natural language processing and information systems - 18th international conference on applications of natural language to information systems* (pp. 278–284).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015a). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications, 42*(1), 259–268.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015b). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications, 42*(4), 2162–2172.

Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence, 26*(1), 25–33.

Rechenthin, M., Street, W. N., & Srinivasan, P. (2013). Stock chatter: using stock sentiment to predict price direction. *Algorithmic Finance, 2*(3), 169–196.

Schumaker, R. P., & Chen, H. (2009a). A quantitative stock prediction system based on financial news. *Information Processing & Management, 45*(5), 571–583.

Schumaker, R. P., & Chen, H. (2009b). Textual analysis of stock market prediction using breaking financial news: the azfin text system. *ACM Transactions on Information Systems, 27*(2), 12:1–12:19.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st annual meeting of the association for computational linguistics, volume 2: short papers* (pp. 24–29). The Association for Computer Linguistics.

Ticknor, J. L. (2013). A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications, 40*(14), 5501–5506.

Tsibouris, G., & Zeidenberg, M. (1995). Testing the efficient markets hypothesis with gradient descent algorithms. In *Neural networks in the capital markets* (pp. 127–136). Wiley: Chichester.

Tumarkin, R., & Whitelaw, R. F. (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal, 57*(3), 41–51.

Vu, T. T., Chang, S., Ha, Q. T., & Collier, N. (2012). An experiment in integrating sentiment features for tech stock prediction in Twitter. In *24th international conference on computational linguistics* (pp. 23–38).

Walczak, S. (2001). An empirical analysis of data requirements for financial forecasting with neural networks. *Journal of Management Information Systems, 17*(4), 203–222.

Xie, B., Passonneau, R. J., Wu, L., & Creamer, G. (2013). Semantic frames to predict stock price movement. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 873–883).

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 - Social and Behavioral Sciences, 26*(0), 55–62.

Zhao, W. X., Jiang, J., Yan, H., & Li, X. (2010). Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 56–65). Association for Computational Linguistics.

Zuo, Y., & Kita, E. (2012a). Stock price forecast using Bayesian network. *Expert Systems with Applications: An International Journal, 39*(8), 6729–6737.

Zuo, Y., & Kita, E. (2012b). Up/down analysis of stock index by using Bayesian network. *Engineering Management Research, 1*(2), 46–52.