Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction

Thien Hai Nguyen **Kiyoaki Shirai**
School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
{nhthien, kshirai}@jaist.ac.jp

Abstract

The goal of this research is to build a model to predict stock price movement using sentiments on social media. A new feature which captures topics and their sentiments simultaneously is introduced in the prediction model. In addition, a new topic model TSLDA is proposed to obtain this feature. Our method outperformed a model using only historical prices by about 6.07% in accuracy. Furthermore, when comparing to other sentiment analysis methods, the accuracy of our method was also better than LDA and JST based methods by 6.43% and 6.07%. The results show that incorporation of the sentiment information from social media can help to improve the stock prediction.

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 large amounts of mood data available. Therefore, incorporating information from social media with the historical prices can improve the predictive ability of the models.

The goal of our research is to develop a model to predict a stock price movement using information from social media (Message Board). In our proposed method, the model predicts the movement of the stock value at t using features derived from information at $t-1$ and $t-2$, where t stands for a transaction date. It 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 movement. That is, the mood of topics in social media is extracted by sentiment analysis. Then, the topics and their sentiments are integrated into the model to predict the stocks. To achieve this goal, discovering the topics and sentiments in a large amount of social media is important to get opinions of investors as well as events of companies. 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 the sentiments from social media have no predictive capabilities (Antweiler and Frank, 2004; Tumarkin and Whitelaw, 2001), while other researchers have reported either weak or strong predictive capabilities (Bollen et al., 2011). Therefore, how to use opinions in social media for stock price predictions is still an open problem.

Our contributions are summarized as follows:

1. We propose a new feature "topic-sentiment" for the stock market prediction model.

2. We propose a new topic model, Topic Sentiment Latent Dirichlet Allocation (TSLDA), which can capture the topic and sentiment simultaneously.

3. Large scale evaluation. Most of the previous researches are limited on predicting for one stock (Bollen et al., 2011; Qian and Rasheed, 2007; Si et al., 2013), and the number of instances (transaction dates) in a test set is rather low such as 14 or 15 instances (Bollen

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1354–1364,
Beijing, China, July 26-31, 2015.

et al., 2011; Vu et al., 2012). With only a few instances in the test set, the conclusion might be insufficient. This is the first research that shows good prediction results on evaluation of many stocks using a test set consisting of many 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 explains our model for sentiment analysis by simultaneously inferring the topic and sentiment in the text. Section 4 describes two kinds of datasets required for stock prediction. Section 5 describes our prediction models and also proposes a novel feature based on the topics and sentiments. Section 6 assesses the results of the experiments. Finally, Section 7 concludes our research.

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 et al., 1969; Fama, 1991), 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 the 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 in some degree (Bollen et al., 2011; Qian and Rasheed, 2007; Vu et al., 2012). Degrees of accuracy at 56% hit rate in the predictions are often reported as satisfying results for stock predictions (Schumaker and Chen, 2009b; Si et al., 2013; Tsibouris and 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 the 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 (Zuo and Kita, 2012a; Zuo and Kita, 2012b). To discover the pattern in the data, they used Bayesian network (Zuo and Kita, 2012a; Zuo and Kita, 2012b), time-series method such as Auto Regressive, Moving Average, Auto Regressive Moving Average model (Zuo and Kita, 2012a) and so on.

2.1 Extracting Opinions from Text

Sentiment analysis has been found to play a significant role in many applications such as product and restaurant reviews (Liu and Zhang, 2012; Pang and Lee, 2008). There are some researches trying to apply sentiment analysis on information sources to improve the stock prediction model. There are two main such sources. In the past, the main source was the news (Schumaker and Chen, 2009a; Schumaker and Chen, 2009b), and in recent years, social media sources. A simple approach is combining the sentiments in 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 tried to use different textual representations such as bag-of-words, noun phrases and named entities for financial news (Schumaker and Chen, 2009b). However, the textual representations are just the words or named entity tags, not exploiting the mood information so much. A novel tree representation based on semantic frame parsers is proposed (Xie et al., 2013). By using stock prices from Yahoo Finance, they annotated all the news in a transaction date with 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 is a task of text classification, not stock prediction.

Naive Bayes was used to classify messages from message boards into three classes: buy, hold and sell (Antweiler and Frank, 2004). They were integrated into the regression model. However, they concluded that their model does not successfully predict stock returns.

A method to measure collective hope and fear on each day and analyze the correlation between these indices and the stock market indicators was

proposed (Zhang et al., 2011). They used the mood words to tag each tweet as fear, worry, hope and so on. They concluded that the ratio of the emotional tweets significantly negatively correlated with Down Jones, NASDAQ and S&P 500, but positively with 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 the positive and negative mood. The latter measures the 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 showed 86.7% direction accuracy (up or down) and 1.79% Mean Absolute Percentage Error. Although they achieved the high accuracy, there were only 15 transaction dates (from December 1 to 19, 2008) in their test set. With such a short period, it might not be sufficient to conclude the effectiveness of their method.

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 was short, from 8^{th} to 26^{th} in September 2012, containing only 14 transaction dates.

Continuous Dirichlet Process Mixture (cDPM) model was used to learn the daily topic set of Twitter messages to predict the stock market (Si et al., 2013). A sentiment time series was built based on these topics. However, the time period of their whole dataset is rather short, only three months.

Most of the researches tried to extract only the opinions or sentiments. However, one important missing thing is that opinions or sentiments are expressed on topics or aspects of companies. Therefore, understanding on which topics of a given stock people are expressing their opinion is very important. Although the models for inferring the topics and sentiments simultaneously have already proposed as discussed in Subsection 2.2, to the best of our knowledge, such models have never applied for stock market prediction.

2.2 Aspect based Sentiment Analysis

Some researches tried to identify the sentiment expressed toward an aspect in a sentence rather than a whole sentence or document. The simple approach is to define a sentiment score of a given aspect by the weighted sum of opinion scores of all words in the sentence, where the weight is defined by the distance from the aspect (Liu and Zhang, 2012; Pang and Lee, 2008). This method is further improved by identifying the aspect-opinion relations using tree kernel method (Nguyen and Shirai, 2015).

Other researches trying to extract both the topic and sentiment for some domains such as online product, restaurant and movie review dataset. ASUM is a model for extracting both the aspect and sentiment for online product review dataset (Jo and Oh, 2011). Joint sentiment/topic model (JST) is another model to detect the sentiment and topic simultaneously, which was applied for movie review dataset (Lin and He, 2009). These models assume that each word is generated from a joint topic and sentiment distribution. It means that these models do not distinguish the topic word and opinion word distributions.

Besides the general opinion words, topic models considering aspect-specific opinion words were also proposed. MaxEnt-LDA hybrid model can jointly discover both aspects and aspect-specific opinion words on a restaurant review dataset (Zhao et al., 2010), while FACTS, CFACTS, FACTS-R, and CFACTS-R model were proposed for sentiment analysis on a product review data (Lakkaraju et al., 2011). However, one of the weaknesses of these methods is that there is only one opinion word distribution corresponding to one topic (aspect). It makes difficult to know which sentiment (e.g. positive or negative) is expressed by the opinion words on that topic.

To overcome this drawback, we propose a new topic model called Topic Sentiment Latent Dirichlet Allocation (TSLDA), which estimates different opinion word distributions for individual sentiment categories for each topic. To the best of our knowledge, such a model has not been proposed. TSLDA is suitable for not only sentiment analysis for stock prediction but also general sentiment analysis of the document, sentence and aspect.

3 TSLDA: Topic Sentiment Latent Dirichlet Allocation

The proposed model TSLDA infers the topics and their sentiments simultaneously. It is an extended model of Latent Dirichlet Allocation (LDA) (Blei et al., 2003). We assume that one sentence ex-

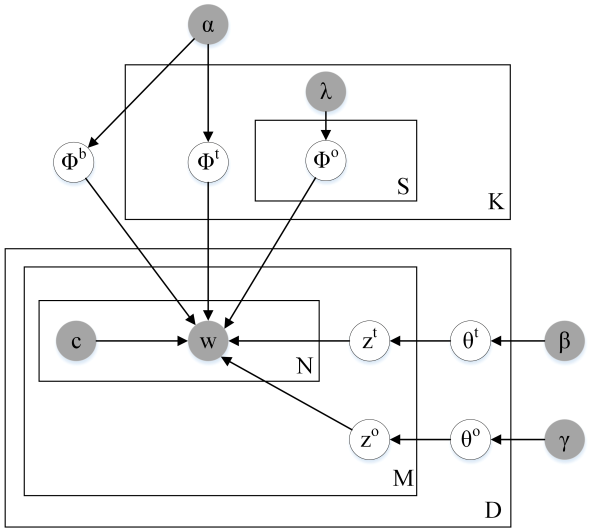

Figure 1: Graphical Model Representation of TSLDA

presses only one topic and one opinion on that topic. The topics are usually nouns, whereas the opinion words are adjectives or adverbs. The words in the document are classified into three categories, the topic word (category $c = 1$), opinion word ($c = 2$) and others ($c = 0$). Then, we suppose the different opinion words are used for the different topics. Depending on the topic, an opinion word may express different sentiment meaning. For example, the opinion word "low" in "low cost" and "low salary" have opposite polarity. In our model, different topics, which are also represented by word distributions, will have different opinion word distributions. Finally, to capture the sentiment meanings such as positive, negative or neutral of the opinion words for each topic, we distinguish opinion word distributions for different sentiment meanings.

Figure 1 shows the graphical model representation of TSLDA. Observed and hidden variables are indicated by shaded and clear circles, respectively. Table 1 shows the notations in Figure 1. The generation process in TSLDA is as follows:

1. Choose a distribution of background words $\Phi^b \sim Dirichlet(\alpha)$
2. For each topic k:
   - Choose a distribution of topic words $\Phi^t_k \sim Dirichlet(\alpha)$
   - For each sentiment s of topic k:
     - Choose a distribution of sentiment words $\Phi^o_{k,s} \sim Dirichlet(\lambda)$

Table 1: Notations in TSLDA

Notation	Definition
$\alpha, \beta, \gamma, \lambda$	Dirichlet prior vectors
K	# of topics
S	# of sentiments
Φ^b	distribution over background words
Φ^t	distribution over topic words
Φ^o	distribution over sentiment words
D	# of documents
M_d	# of sentences in document d
$N_{d,m}$	# of words in sentence m in document d
θ^t_d	topic distribution for document d
θ^o_d	sentiment distribution for document d
$z^t_{d,m}$	topic assignment for sentence m in document d
$z^o_{d,m}$	sentiment assignment for sentence m in document d
$w_{d,m,n}$	n^{th} word in sentence m in document d
$c_{d,m,n}$	n^{th} word's category (background, topic or sentiment) in sentence m in document d

3. For each document d:
   - Choose a topic distribution $\theta^t_d \sim Dirichlet(\beta)$
   - Choose a sentiment distribution $\theta^o_d \sim Dirichlet(\gamma)$
   - For each sentence m:
     - Choose a topic assignment $z^t_{d,m} \sim Multinomial(\theta^t_d)$
     - Choose a sentiment assignment $z^o_{d,m} \sim Multinomial(\theta^o_d)$
     - For each word in the sentence:
       - Choose a word $w_{d,m,n}$ as in Equation (1).

$$w_{d,m,n} \sim \begin{cases} Multinomial(\Phi^b) & \text{if } c_{d,m,n} = 0 \\ Multinomial(\Phi^t_{z^t_{d,m}}) & \text{if } c_{d,m,n} = 1 \\ Multinomial(\Phi^o_{z^t_{d,m}, z^o_{d,m}}) & \text{if } c_{d,m,n} = 2 \end{cases} \quad (1)$$

We will define some notations for explanation of our method. $W^{k,s}_{d,m,v,c}$ is the number of times the word v with the category c appears in the sentence m in the document d, where m discusses the topic k and the sentiment s. Let $Z^{k,s}_d$ be the number of times the document d has the topic k and the sentiment s. If any of these dimensions is not limited

to a specific value, we used an asterisk $*$ to denote it. For example, $W^{k,s}_{*,*,v,c}$ is the number of appearance of combination (v, c, k, s) in any sentences in any documents. Similarly, $Z^{k,*}_d$ is the number of times the document d has the topic k with any sentiments.

A bold-font variable denotes the list of the variables. For instance, $\boldsymbol{z^t}$ and $\boldsymbol{w}$ denote all of topic assignments and words in all documents, respectively.

$-(d, m)$ stands for exclusion of the value in the sentence m in the document d. For example, $\boldsymbol{z^t_{-(d,m)}}$ denotes all of topic assignment variables $\boldsymbol{z^t}$ but $z^t_{d,m}$. $Z^{a,*-(d,m)}_d$ denotes the value of $Z^{a,*}_d$ not counting times at the sentence m in the document d.

We used square brackets for specifying the value at the index of a vector or distribution. For instance, $\alpha[v]$ denotes the value of α at index v.

Collapsed Gibbs Sampling was implemented for inference in TSLDA. It will sequentially sample hidden variables $z^t_{d,m}$ and $z^o_{d,m}$ from the distribution over these variables given the current values of all other hidden and observed variables. In other words, in order to perform Collapsed Gibbs Sampling, conditional probability $P(z^t_{d,m} = a, z^o_{d,m} = b|\boldsymbol{z^t_{-(d,m)}}, \boldsymbol{z^o_{-(d,m)}}, \boldsymbol{w}, \boldsymbol{c})$ is calculated by marginalizing out random variables Φ^b, $\boldsymbol{\Phi^t}$, $\boldsymbol{\Phi^o}$, $\boldsymbol{\theta^t}$ and $\boldsymbol{\theta^o}$. Because of the limit of spaces, we only show the final formula of this conditional probability as in Equation (2). Let $V_{d,m}$ be a set of words in the sentence m in the document d. V is a set of all of the words in all documents.

$$
\begin{aligned}
&P(z^t_{d,m} = a, z^o_{d,m} = b|\boldsymbol{z^t_{-(d,m)}}, \boldsymbol{z^o_{-(d,m)}}, \boldsymbol{w}, \boldsymbol{c},) \\
&\propto (Z^{a,*-(d,m)}_d + \beta[a])(Z^{*,b-(d,m)}_d + \gamma[b]) \\
&\times \frac{\prod_{v=1}^{V_{d,m}} \prod_{j=1}^{W^{*,*}_{d,m,v,1}} (W^{a,*-(d,m)}_{*,*,v,1} + \alpha[v] + j - 1)}{\prod_{j=1}^{W^{*,*}_{d,m,*,1}} (\sum_{v=1}^{V} W^{a,*-(d,m)}_{*,*,v,1} + \alpha[v] + j - 1)} \\
&\times \frac{\prod_{v=1}^{V_{d,m}} \prod_{j=1}^{W^{*,*}_{d,m,v,2}} (W^{a,b-(d,m)}_{*,*,v,2} + \lambda[v] + j - 1)}{\prod_{j=1}^{W^{*,*}_{d,m,*,2}} (\sum_{v=1}^{V} W^{a,b-(d,m)}_{*,*,v,2} + \lambda[v] + j - 1)}
\end{aligned}
\tag{2}
$$

Multinomial parameters: Finally, samples obtained from Collapsed Gibbs Sampling can be used to approximate the multinomial parameter sets. The distributions of topics and sentiments in the document d are estimated as in Equation (3).

$$
\theta^t_d[a] = \frac{Z^{a,*}_d + \beta[a]}{\sum_{k=1}^{K} Z^{k,*}_d + \beta[k]};\ \theta^o_d[b] = \frac{Z^{*,b}_d + \gamma[b]}{\sum_{s=1}^{S} Z^{*,s}_d + \gamma[s]} \tag{3}
$$

The background word distribution, topic word distribution of the topic k and sentiment word distribution of the sentiment s for k are estimated in Equation (4), (5) and (6), respectively.

$$
\Phi^b[r] = \frac{W^{*,*}_{*,*,r,0} + \alpha[r]}{\sum_{v=1}^{V} W^{*,*}_{*,*,v,0} + \alpha[v]} \tag{4}
$$

$$
\Phi^t_k[r] = \frac{W^{k,*}_{*,*,v,1} + \alpha[r]}{\sum_{v=1}^{V} W^{k,*}_{*,*,v,1} + \alpha[v]} \tag{5}
$$

$$
\Phi^o_{k,s}[r] = \frac{W^{k,s}_{*,*,v,2} + \lambda[r]}{\sum_{v=1}^{V} W^{k,s}_{*,*,v,2} + \lambda[v]} \tag{6}
$$

4 Dataset

Two datasets are used for the development of our stock prediction model. One is the historical price dataset, and the other is the message board dataset.

4.1 Historical Price Dataset

Historical prices are extracted from Yahoo Finance for 5 stocks. The list of the stock quotes and company names is shown in Table 2. 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. They are 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.

4.2 Message Board Dataset

To get the mood information of the stocks, we collected 5 message boards of the 5 stocks from Yahoo Finance Message Board for a period of one year (from July 23, 2012 to July 19, 2013). On the message boards, users usually discuss company

Table 2: Statistics of Our Dataset

Stocks	Company Names	#Documents
XOM	Exxon Mobil Corporation	11027
DELL	Dell Inc.	10339
EBAY	eBay Inc.	7168
IBM	International Business Machines Corporation	5008
KO	The Coca-Cola Company	2024

news, prediction about stock going up or down, facts, comments (usually negative) about specific company executives or company events. 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 it is the time of closing transaction. There are 249 transaction dates in the one year period in our dataset.

5 Stock Prediction Models with Sentiment Analysis

This paper focuses on prediction of not the stock price but movement of it. That is, our goal is to develop a model that predicts if the stock price goes up or down. 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 many tasks such as text classification (Joachims, 1998; Nguyen and Shirai, 2013). Therefore, we chose SVM with the linear kernel as the prediction model. Furthermore, features derived by sentiment analysis on the message board are incorporated in it. To assess the effectiveness of sentiment analysis, four sets of features are designed. The first one uses only the historical prices. The other sets include topic and sentiment features obtained by different methods. All the feature values are scaled into $[-1, 1]$ value. Table 3 summarizes our features used in the model to predict the price movement at the transaction date t. The details of each feature will be explained in the next subsections.

5.1 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 of the price movement in the history of the stock. In addition, it is a baseline for evaluation of the effectiveness of the sentiment features. Features used for training SVM are $price_{t-1}$ and $price_{t-2}$ which are the price movements (up, down) at the transaction dates $t-1$, $t-2$, respectively.

Table 3: Features of the Prediction Model

Method	Features
Price Only	$price_{t-1}, price_{t-2}$
LDA-based Method	$price_{t-1}, price_{t-2}$, $lda_{i,t}, lda_{i,t-1}$
JST-based Method	$price_{t-1}, price_{t-2}$, $jst_{i,j,t}, jst_{i,j,t-1}$
TSLDA-based Method	$price_{t-1}, price_{t-2}$, $tslda_{i,j,t}, tslda_{i,j,t-1}$

5.2 LDA-based Method

In this model, we consider each message as a mixture of hidden topics. LDA is a generative probabilistic model of a corpus [1]. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Hidden topics of LDA are incorporated into the prediction model as follows. First, stop words are removed from the messages, and all the words are lemmatized by Stanford CoreNLP (Manning et al., 2014). Topics are inferred by Gibbs Sampling with 1000 iterations. Next, the probability of each topic for each message is calculated. For each transaction date t, the probability of each topic is defined as the average of the probabilities of the topic in all messages posted on that transaction date.

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 ($i \in \{1, \cdots, K\}$) for the transaction dates t and $t-1$. The number of the topics K is empirically determined as explained in Subsection 6.1.

5.3 JST-based Method

When people post the message on social media to express their opinion for a given stock, they tend to talk their opinions for a given topic or aspect such as profit and dividend. They would think that the future price of the stock goes up or down by seeing pairs of topic-sentiment written by others. Following the above intuition, we propose a new feature topic-sentiment for the stock predic-

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

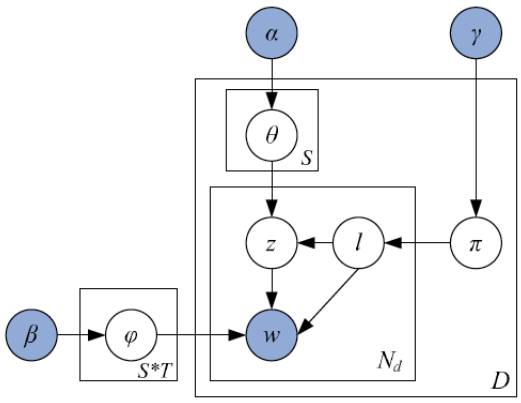

Figure 2: Graphical Model Representation of JST

Table 4: Notations in JST

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

tion model. Two methods are used to extract the pairs of topic-sentiment from the message board. One is a latent topic based model called JST (Lin and He, 2009). The other is TSLDA discussed in Section 3. This subsection introduces the method using the former.

We consider each message as a mixture of hidden topics and sentiments. JST model is used to extract topics and sentiments simultaneously. Figure 2 shows the graphical model representation of JST. Notations in Figure 2 are shown in Table 4. In LDA model, there is only one document specific topic distribution. In contrast, each document in JST is associated with multiple sentiment labels. Each sentiment label is associated with a document specific topic distribution. A word in the document is drawn from a distribution over words defined by the topic and sentiment label.

After removal of stop words and lemmatization, JST model is trained by Gibbs Sampling with 1000 iterations. We chose 3 as the number of sentiments which might represent negative, neutral and positive. The number of the topics K is empirically determined as explained in Subsection 6.1. Next, the joint probability of each pair of topic and sentiment is calculated for each message. For each transaction date t, the joint probability of each topic-sentiment pair is defined as the average of the joint probabilities in the messages on that transaction date. Then we integrate 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 ($j \in \{1, \cdots, K\}$) for the transaction dates t and $t-1$.

5.4 TSLDA-based Method

We use our TSLDA model to capture the topics and sentiments simultaneously. First, a rule-based algorithm is applied to identify the category of each word in the documents. Consecutive nouns are considered as topic words. If a word is not a noun and in a list of opinion words in SentiWordNet (Baccianella et al., 2010), it is considered as an opinion word. The rest of words are classified as background words.

After lemmatization, TSLDA model is trained by Collapsed Gibbs Sampling with 1000 iterations. We chose 3 as the number of sentiments which might represent for negative, neutral and positive. K (number of topics) is determined as explained in Subsection 6.1. The topic and its sentiment in each sentence are gotten from the topic assignment and sentiment assignment in TSLDA. If there is a sentence expressing the sentiment j on the topic i, we represent the tuple $(i, j) = 1$, and 0 otherwise. The proportion of (i, j) over all sentences are calculated for each message. For each transaction date, a weight of the tuple (i, j) is defined as the average of the proportions over all messages. Then we integrated the weights of the topics and their sentiments into the prediction model.

Features used for training SVM are $price_{t-1}$, $price_{t-2}$, $tslda_{i,j,t}$ and $tslda_{i,j,t-1}$. $tslda_{i,j,t}$ and $tslda_{i,j,t-1}$ are the weights of the topic i ($i \in \{1, \cdots, K\}$) with the sentiment j ($j \in \{1, 2, 3\}$) for the transaction dates t and $t-1$.

Table 5: Accuracies of Stock Movement Prediction

Stocks	Price Only	LDA	JST	TSLDA
XOM	0.5000	0.4464	0.5179	**0.5357**
DELL	**0.5893**	0.5357	0.5000	0.5536
EBAY	0.6071	0.6071	0.5000	**0.6429**
IBM	0.4107	0.3929	0.5357	**0.5536**
KO	0.4107	0.5179	0.4643	**0.5357**
Average	0.5036	0.5000	0.5036	**0.5643**

Table 6: TP, TN, FP, FN of Stock Movement Prediction

Stocks	Metrics	Price Only	LDA	JST	TSLDA
XOM	TP	14	13	15	18
	TN	14	12	14	12
	FP	8	10	8	10
	FN	20	21	19	16
DELL	TP	17	13	5	13
	TN	16	17	23	18
	FP	17	16	10	15
	FN	6	10	18	10
EBAY	TP	17	18	20	20
	TN	17	16	8	16
	FP	9	10	18	10
	FN	13	12	10	10
IBM	TP	15	15	7	31
	TN	8	7	23	0
	FP	17	18	2	25
	FN	16	16	24	0
KO	TP	12	14	16	10
	TN	11	15	10	20
	FP	17	13	18	8
	FN	16	14	12	18
Sum	TP	75	73	63	92
	TN	66	67	78	66
	FP	68	67	56	68
	FN	71	73	83	54

6 Evaluation

6.1 Experiment Setup

We divided the dataset described in Section 4 into three parts: training set from July 23, 2012 to March 31, 2013, development set from April 01, 2013 to April 30, 2013, and test set from May 01, 2013 to July 19, 2013. The label of 'up' and 'down' is assigned to each transaction date by comparing the price of the current and previous dates.

To optimize the number of topics K for each stock, we run the models with four values of K: 10, 20, 50 and 100. The best K is chosen for each stock on the development set, and the systems with the chosen K is evaluated on the test data. The performance of the prediction is measured by accuracy.

For the hyperparameters of LDA, JST and TSLDA, we simply selected symmetric Dirichlet prior vectors, that is all possible distributions are likely equal. We used the default values of these hyperparameters for LDA and JST. Concretely speaking, $\alpha = 0.5$, $\beta = 0.01$ in LDA and $\alpha = \frac{50}{\#topics}$, $\beta = 0.01$, $\gamma = 0.3$ were used in JST. For TSLDA, we set $\alpha = 0.1$, $\lambda = 0.1$, $\beta = 0.01$ and $\gamma = 0.01$.

6.2 Results

The result of each stock is shown in Table 5. In addition, the average of 5 stocks for each model is revealed in the last row of this table for easy comparison. Our model TSLDA-based method outperformed the other methods on the average of the stocks. Table 6 shows the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) of models for the stocks. For easy comparison, the summation for these five stocks are calculated in the last row.

To assess the effectiveness of integrating mood information, we compare our TSLDA-based method with Price Only method. The results showed that the model using mood information outperformed the model without mood by 3.57%, 3.58%, 14.29% and 12.5% accuracy for XOM, EBAY, IBM and KO stock, respectively. On the other hand, the performance on DELL stock was not improved. It means that the use of the mood does not always make the performance better. The mood from social media could lead to a wrong prediction because of wrong prediction of message writers, fault information and so on. However, TSLDA was better than Price Only method on average of these stocks. In addition, TSLDA can reduce the number of FN, especially for IBM, although FP was not changed in the sum of 5 stocks. Thus, we can conclude that integrating the mood information from social media can help to predict stock price movement more precisely.

Next, let us compare the models for inferring latent topics only (LDA) and topics and sentiments (JST and TSLDA) in the stock movement prediction. The accuracy of JST-based method was better than LDA for two stocks (XOM and IBM), worse for three stocks and comparable in the average of five stocks. While, TSLDA-based method outperformed LDA and JST by 2 to 17% in the accuracy for five stocks. TSLDA was also better

Table 7: Top Words in Topics of TSLDA

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
ko	split	drink	customer	company	country
ceo	stock	coke	budget	competitor	tax
company	share	water	campaign	buy	governor
report	price	produce	promotion	sell	obama
earning	dividend	product	growth	hold	rommey
analyst	year	health	sale	problem	mitt
share	date	juice	volumn	soda	president
news	market	make	come	product	bill
downgrade	time	p.o.s	revenue	people	christian

Table 8: Top Words in Sentiments of Topics of TSLDA

Topic1			Topic2		
S1	S2	S3	S1	S2	S3
old	value	grow	down	straight	good
tired	even	strong	tough	warm	long
unreal	difference	solid	troll	informative	more
much	list	gain	breakthrough	interesting	high
obviously	together	full	ex	later	still
much	serve	continue	sugary	responsible	right
not	americans	growth	ep	yeah	sure
helpful	operation	value	richly	used	same
here	get	quarter	major	though	many

than LDA and JST on average as shown in Table 5. The improvement of the accuracy was derived by increase of TP and decrease of FN. These results indicate that (1) our idea to use both latent topics and sentiments as the features is effective, (2) TSLDA is more appropriate model than JST in stock movement prediction.

Table 7 shows examples of highly associated words of some topics for stock KO (Coca-Cola Company) in TSLDA. For example, 'split', 'stock' and 'share' are words highly associated with the hidden topic 2, and 'drink', 'coke' and 'water' are highly associated with the topic 3. The first five hidden topics in Table 7 may represent the management, stock market trading, product, customer care service, competitors of the company, while the last one indicates macroeconomic factors. Table 8 shows examples of highly associated words of three sentiments of the hidden topic 1 and 2. For the hidden topic 1, 'growth', 'strong', 'solid' etc. are the words highly associated with the hidden sentiment 3 (which may corresponds to positive class), while 'old', 'tired', 'unreal' etc. with the hidden sentiment 1 (may be negative). In general, however, it is rather difficult to interpret the meaning of the hidden sentiment because the sentiments have many dimensions such as happy, anger, sad, vital and so on. We also found that the words with high probabilities in the background distribution were the stop words, punctuations, function words, messy characters written in social media, e.g. '.', 'the', 'and', 'you', '$', 'for' and '?'.

Table 9 shows top words in some joint sentiment topic distributions of JST model for stock KO. For example, 'yahoo', 'ko' and 'finance' are highly associated with the distribution defined by hidden sentiment 1 and hidden topic 1. However, it is rather difficult to guess which sentiment or topic in this joint distribution actually means.

Table 9: Top Words in Distributions Defined by Sentiments and Topics of JST

S1		S2		S3	
Topic1	Topic2	Topic1	Topic2	Topic1	Topic2
yahoo	juice	ko	new	spam	split
ko	minute	buy	american	board	share
finance	maid	get	country	post	date
chart	orange	sell	obama	ignore	stock
free	apple	go	top	idiot	record
fire	drink	make	fall	get	price
website	fruit	money	health	read	august
aone	edit	much	government	another	receive
download	punch	next	place	report	get

7 Conclusion

This paper presents the method to infer the topics and their sentiments on the documents and use them for prediction of the stock movement. The results of the experiments show the effectiveness of our proposed TSLDA-based method. Although 56% accuracy of our method is not so high, it can be satisfying results as regarded in the previous papers. Another advantage of the paper is the evaluation by the large scale experiment (five stocks, three month transaction dates in the test set).

The drawback of TSLDA is that we have to specify the number of topics and sentiment beforehand. To overcome it, TSLDA should be extended as a non-parametric topic model estimating the number of topics inherent in the data. This will be done in our future work.

References

Werner Antweiler and Murray Z Frank. 2004. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 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 (LREC'10), volume 10, pages 2200–2204.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Eugene F Fama, Lawrence Fisher, Michael C Jensen, and Richard Roll. 1969. The adjustment of stock prices to new information. *International economic review*, 10(1):1–21.

Eugene F Fama. 1991. Efficient capital markets: Ii. *The journal of finance*, 46(5):1575–1617.

Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824. ACM.

Thorsten Joachims. 1998. *Text categorization with support vector machines: Learning with many relevant features*. Springer.

Himabindu Lakkaraju, Chiranjib Bhattacharyya, Indrajit Bhattacharya, and Srujana Merugu. 2011. Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In *Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, USA*, pages 498–509. SIAM / Omnipress.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384. ACM.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pages 415–463. Springer.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Thien Hai Nguyen and Kiyoaki Shirai. 2013. Text classification of technical papers based on text segmentation. In Elisabeth Mtais, Farid Meziane, Mohamad Saraee, Vijayan Sugumaran, and Sunil Vadera, editors, *Natural Language Processing and Information Systems*, volume 7934 of *Lecture Notes in Computer Science*, pages 278–284. Springer Berlin Heidelberg.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Aspect-based sentiment analysis using tree kernel based relation extraction. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 9042 of *Lecture Notes in Computer Science*, pages 114–125. Springer International Publishing.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Bo Qian and Khaled Rasheed. 2007. Stock market prediction with multiple classifiers. *Applied Intelligence*, 26(1):25–33.

Michael Rechenthin, W Nick Street, and Padmini Srinivasan. 2013. Stock chatter: Using stock sentiment to predict price direction. *Algorithmic Finance*, 2(3):169–196.

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

Robert P. Schumaker and Hsinchun Chen. 2009b. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, 27(2):12:1–12:19, March.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. 2013. Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 24–29. The Association for Computer Linguistics.

George Tsibouris and Matthew Zeidenberg. 1995. Testing the efficient markets hypothesis with gradient descent algorithms. In *Neural Networks in the Capital Markets*, pages 127–136. Wiley: Chichester.

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

Tien Thanh Vu, Shu Chang, Quang Thuy Ha, and Nigel Collier. 2012. An experiment in integrating sentiment features for tech stock prediction in twitter. In *24th International Conference on Computational Linguistics*, pages 23–38.

Steven Walczak. 2001. An empirical analysis of data requirements for financial forecasting with neural networks. *Journal of management information systems*, 17(4):203–222.

Boyi Xie, Rebecca J Passonneau, Leon Wu, and Germán Creamer. 2013. Semantic frames to predict stock price movement. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 873–883.

Xue Zhang, Hauke Fuehres, and Peter A Gloor. 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.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 56–65. Association for Computational Linguistics.

Yi Zuo and Eisuke Kita. 2012a. Stock price forecast using bayesian network. *Expert Systems with Applications: An International Journal*, 39(8):6729–6737.

Yi Zuo and Eisuke Kita. 2012b. Up/down analysis of stock index by using bayesian network. *Engineering Management Research*, 1(2):46–52.