Exploiting Social Media for Stock Market Prediction with Factorization Machine

Chen Chen[1], Wu Dongxing[1], Hou Chunyan[2,*], Yuan Xiaojie[1]

[1]College of Computer and Control Engineering
Nankai University
Tianjin, China
{nkchenchen,yuanxj}@nankai.edu.cn,
wudongxing@dbis.nankai.edu.cn

[2]School of Computer and Communication Engineering
Tianjin University of Technology
Tianjin, China
houchunyan@tjut.edu.cn

***Abstract*—When the stock market has become more and more competitive, the stock market prediction has been a hot research topic. Traditional methods are based on historical stock data, which ignore the latest market information. Later although financial news is proposed to access market information, there are some disadvantages for news to predict the stock market. Recently when the micro-blogging service has grown to a popular social media and provides a number of real-time messages for a lot of users, social media is proposed for the stock market prediction. In that case, the high-dimension of textual feature poses a main challenge. In this study, we propose a novel kind of model, Factorization Machine (FM), to predict the trend of stock market. FM not only alleviates the impact of high dimensionality, but also captures some aspects of basic linguistics. Experiments on real-world data show that FM can achieve 81% accuracy and get significantly more profit than state-of-the-art models. In addition, we shed light on how textual representation influences the prediction and find that FM is stable, which is applicable to other social media or prediction application generally.**

Keywords—social media; stock market prediction; factorization machines; microblog

I. INTRODUCTION

Since the stock market is an important and active part of financial markets, the stock market prediction has achieved widespread concern from academic and business communities. There are several economic theories about future prices of stock market. The first is Efficient Market Hypothesis [9], which asserts that the price reflects all known information. Another is Random Walk Hypothesis [22], which believes that the price movement is random. These theories mean that the price of stock cannot be predicted effectively. However, recent researches suggest that stock prices could be predicted to some extent from the perspective of socioeconomic theory of finance and behavioral economics [11, 25, 28].

Traditionally, the stock market prediction is based on historical stock data. The various technical indicators can be extracted from historical data to predict the trend of stock market [5, 7, 18, 27]. These studies focus on the past market information, but ignore the latest information.

Financial news articles, known as one of the most important part of market information, are used as the latest information for the stock market prediction [10, 19, 35, 39]. This kind of news includes big mergers, bankruptcy of some companies or economic crisis. Financial news usually has significant impact on stock market because traders rely on them to make judgment for future trading decisions. However, financial news articles have following disadvantages. Firstly, financial news just reflects the viewpoints of editors and reporters. Secondly, because the news release is not scheduled, it is necessary to align news articles with the time of stock market. Thirdly, web pages of financial news are full of a large amount of advertisement and consist of more noise data.

Recently, with the rise of social media like Facebook and Twitter[1] in USA, Sina Weibo[2] in China, some researchers try to extract information from social media to predict changes in various economic and commercial indicators. For example, Gruhl *et al.* [13] show how online chat activity predicts book sales. Mishne and Glance [23] use assessments of blog sentiment to predict movie sales. Liu *et al.* [20] analyze the sentiment from blogs to predict future product sales. In addition, Google search queries have been shown to provide early indicators of disease infection rates and consumer spending [6]. Inspired by these studies, some researchers explore social media for stock market prediction, assuming that the sentiment or emotion is one of the vital factors that can influence the stock market [3, 12, 26]. For example, Johan Bollen *et al.* [3] find that micro blogs that are labeled as "calm" have the powerful prediction ability to the Dow Jones industries average index with highest accuracy over 80%. However, they need to analyze collective mood states derived from large-scale textual messages of Twitter feeds.

To carry out accurate prediction based on textual messages, researchers have tried various models and algorithms, and have achieved considerable results. Schumaker and Chen [35] survey a variety of machine learning algorithms. When predicting the future prices of stock market based on textual data, the high dimensionality is a great challenge. As methods of dimensionality reduction, factorization models are used for

[1] http://twitter.com
[2] http://weibo.com

978-1-4799-4143-8/14 $31.00 © 2014 IEEE
DOI 10.1109/WI-IAT.2014.91

many kinds of application [30, 31]. However, little prior research work has been done to handle the problem of stock market prediction using factorization model. This is partially due to the fact that, for short term stock market, the real world dataset is not in large scale enough to make the factorization model effective. Social media can provide sufficient data due to timely updates and the intensive interaction by the large number of user. In this case, factorization models not only are used for dimensionality reduction, but also capture some aspects of basic linguistic, such as the synonymy and collocation, to improve the prediction. Our experiments show that factorization models are significantly better than several non-trivial models.

In this paper, we start to investigate how to exploit social media for stock market prediction with factorization machine. To the best of our knowledge, we are the first group of researchers to elaborate stock market prediction with factorization machine. Our contributions are three-fold:

1. Analyze the relationship of factorization machine (FM) to Generalized Linear Model (GLM) and Support Vector Model (SVM), and provide insights on why FM is better than GLM and SVM for high-dimensional data from the perspective of model formulation.

2. Exploit social media for stock market prediction on explicit textual features. Explain the advantage of social media for stock market prediction in contrast to other data source.

3. Demonstrate the superior performance of FM in the stock market prediction by comparing with several non-trivial baselines. Investigate the sensitivity of factorization machine to the textual representation.

The rest of the paper is organized as follows. In section 2, we summarize the related work about stock market prediction. In section 3, we introduce the predictive model, that is, factorization machine, and derive its relation to generalized linear model and support vector model. In section 4, we evaluate proposed model against a number of non-trivial baselines. We conclude our paper in section 5.

II. RELATED WORK

In this section, we review three lines of relevant research work: 1) stock market prediction, 2) prediction based on social media, 3) application of factorization machine. We associate them with our work and discuss the novelty of our study as well.

A. Stock Market Prediction

Historical data of stock market are the popular source of market information. The historical index data transformed into various technical indicators as fundamental factors are considered as inputs to the proposed models [7, 18, 21, 34, 36]. Shen *et al.* select some technical indicators with great influence to stock market and train models to predict the trend of Shanghai Composite Index [36]. Dai *et al.* establish model based on selected four technical indicators [7]. However, their works consider past information rather than the latest information which has significant impacts on the behavior of the stock market. We focus on timely and latest information to predict stock market.

Financial news articles, known as one of the most important part of market information, are widely used. With the development of internet, both the speed of news broadcasting and the amount of news articles have been changing. Some studies [10, 24, 35] focus on predicting short-term trend of stock market based on financial news articles. Schumaker and Chen [35] examine a predictive machine learning approach for financial news articles analysis using several different textual representations. However, there are some disadvantages for financial news to predict stock market. We apply a new kind of information source, social media, to predict stock market.

Some researchers focus on sentiment analysis of the text with assumption that the sentiment or emotion is one of the vital factors that can influence the stock market. Johan Bollen *et al.* [3] find that even though general sentiments, such as positive vs. negative, is not helpful in the prediction of qualitative changes in closing values of the Dow Jones Industrial Average (DJIA) and S&P 500, some specific and more detailed sentiments, such as calm, happiness and Anxiety Index, have a predictive power to inform broad direction of stock market in near future. They reveal that micro blogs which are labeled as “calm” have the powerful prediction ability to the Dow Jones industries average index with highest accuracy over 80%. Our work differs from these methods in two ways. Firstly, we do not need to recognize collective mood states derived from large-scale microblog feeds. Secondly, we focus on stock market prediction directly by means of collecting high relevant information timely and analyzing the level of attentions in social media.

B. Prediction Using Social Media

With the launch of Twitter in 2007, micro-blogging service becomes highly popular and a number of researches try to exploit social media for a variety of prediction such as the movie box-office, political election and marketing. Asur and Huberman [1] demonstrate how social media content can be used to predict box-office revenues for movies. Bothos *et al.* have tried to predict the Oscars winners using social media [4]. The number of related social media contents is a valid predictor for successful election [38]. Jansen *et al.* investigate micro-blogging as a form of electronic word-of-mouth for sharing consumer opinions concerning brands [15]. Compared with these studies, we focus on stock market prediction using social media.

C. Application of Factorization Machine

Recently, factorization machines, presented by Rendle [31], are used in some important application and have shown excellent prediction capabilities. Hong *et al.* address the problem of predicting user decisions and modeling users’ interests with factorization machine [14]. Rendle et al. propose to apply factorization machines to model contextual information and to provide context-aware recommendation [32]. Factorization Machine has been proposed for practical applies application in the field of information retrieval [29]. Compared with these applications, we apply the factorization machine to the stock market prediction.

III. FACTORIZATION MACHINE

A. Preliminary

Proposed by [31], Factorization Machine is a state-of-the-art framework for factorization model with a variety of features. It is able to model all nested interactions up to order d between the p input variables in x using factorized interaction parameters. The model of order $d = 2$ is defined as:

$$\hat{y}(x) = w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j=i+1}^{p} \langle v_i, v_j \rangle x_i x_j \tag{1}$$

where k is the dimensionality of the factorization and the model parameter $\Theta = \{w_0, w, V\}$ is:

$$w_0 \in \mathbb{R}, \qquad w \in \mathbb{R}^n, \qquad V \in \mathbb{R}^{n \times k} \tag{2}$$

In addition, $\langle \cdot, \cdot \rangle$ is the dot product of two k-dimension vectors:

$$\langle v_i, v_j \rangle = \sum_{f=1}^{k} v_{i,f} v_{j,f} \tag{3}$$

A row vector v_i of V represents the ith variable with k dimension. The first part of FM model, that is w_0, is the global bias, and the second part contains the unary interactions of each input variable x_i with the target, while the last part with two nested sums consists of all pairwise interactions of input variables x_i and x_j. The weight of the ith variable is measured by w_i, and $\langle v_i, v_j \rangle$ models the pairwise interaction between x_i and x_j. The important characteristic of FM is that the effect of the pairwise interaction is not modeled by an independent parameter $w_{i,j}$ but with a factorized parameter $\langle v_i, v_j \rangle$, which assumes that the effect of pairwise interactions has a low rank. This allows FM to estimate reliable parameters of high order interactions even in highly sparse data.

Rendle [30] proves that FM can be computed efficiently with the computational complexity of $O(k \times p)$ as it is equivalent to:

$$\hat{y}(x) = w_0 + \sum_{i=1}^{p} w_i x_i + \frac{1}{2} \sum_{f=1}^{k} \left(\left(\sum_{i=1}^{p} v_{i,f} x_i \right)^2 - \sum_{i=1}^{p} v_{i,f}^2 x_i^2 \right) \tag{4}$$

FM can mimic the structure of many state-of-the-art model like matrix factorization [37], pairwise interaction tensor factorization [33] and neighborhood models [16] in one unified framework.

In brief, modeling pairwise interaction and factorization parameterization are two main advantages of FM, which play important role in improving the stock market prediction. There are a lot of messages related with stock market in social media each day. Stock market is discussed in these messages from different perspectives such as trading volume, closing price. On the other hand, some verbs, such as rise or drop, are used to express the opinion of uses about the trend of stock market. Thus, the pairwise interactions, rather than single interaction, are able to indicate the trend in linguistics. The factorization parameterization is to project the feature into a smaller dimensional space so that the dot product in the smaller space can measure the strength of pairwise interaction exactly. The large number of textual feature includes many irrelevant and noisy features. Factorization parameterization can reduce the impact of those features.

B. Relation to Generalized Linear Model

To derive a Generalized Linear Model (GLM), the following three assumptions are made about the model:

1. $y \sim ExponentialFamily(\eta)$ .That is, given x and θ, the distribution of y follows some exponential family distribution with parameter η.

2. Given x, the goal is to predict the expected value $E[T(y)]$ of $T(y)$ given x.

3. The natural parameter η and the inputs x are related linearly, as Eq.(5) shown.

In GLM, y is distributed as the exponential family. The distribution of the exponential family is formulized as:

$$p(y;\eta) = p(y) \exp(\eta^T T(y) - a(\eta)) \tag{5}$$

Here, η is called the natural parameter of the distribution, $T(y)$ is the sufficient statistic, and $a(\eta)$ is the log partition function.

In a classification or regression problem, we would like to predict the value of some random variable y as a function of x. An important assumption of GLM is that natural parameter η and the inputs x are related linearly, as shown in Eq. (6).

$$\eta = \sum_{i=1}^{p} w_i x_i \tag{6}$$

If the natural parameter is expressed in Eq.(7), we generalize FM in different applications of GLM.

$$\eta = w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j=i+1}^{p} \langle v_i, v_j \rangle x_i x_j \tag{7}$$

The regression can be modeled by the Gaussian distribution. This model is linear regression model in Eq.(8).

$$a(\eta) = \frac{1}{2} \eta^2, \qquad E(T(y)) = \eta \tag{8}$$

The binary classification can be modeled by the Bernoulli distribution. This model is logistic regression model in Eq.(9).

$$a(\eta) = \log(1 + \exp(\eta)), \qquad E[T(y)] = \frac{\exp(\eta)}{1 + \exp(\eta)} \tag{9}$$

The multi-class problem can be modeled by the multinomial distribution. This model is softmax regression model in Eq.(10).

$$a(\eta) = \log\left(\frac{1}{\sum_{i=1}^{K} \exp(\eta_i)}\right), \qquad E(T(y)_i) = \frac{\exp(\eta_i)}{\sum_{j=1}^{K} \exp(\eta_j)} \tag{10}$$

In summary, GLM assumes that natural parameter η and the inputs x are related linearly while FM uses natural parameter η to express input x and the nested interactions as well. For stock market prediction based on social media, the nested interactions between textual features can capture some aspects of basic linguistics. Thus, from the perspective of model formulation, FM is more likely to produce better prediction than GLM.

C. Relation to Support Vector Machine

The model of SVM can be written as the dot product between the transformed input x and model parameters w:

$$y(x) = \langle \phi(x), w \rangle \tag{11}$$

where $\phi()$ is a feature map from the feature space into Reproducing Kernel Hilbert Space (RKHS). The feature map is related to the kernel:

$$K(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle \tag{12}$$

Rendle [30] discusses the relationship between FM and SVM by different kernel function.

1) Linear kernel: the most simple kernel is the linear kernel: $K(x_1, x_2) = 1 + \langle x_1, x_2 \rangle$, which is associated with the feature map $\phi(x) = (1, x_1, \cdots, x_p)$. Thus, the model of the linear SVM can be expressed as:

$$y(x) = w_0 + \sum_{i=1}^{p} w_i x_i \tag{13}$$

As we can see, a linear SVM is same to a FM of $d = 1$.

2) Polynomial kernel: the polynomial kernel allows the SVM to model higher interactions between variables. It is defined as $K(x_1, x_2) = (\langle x_1, x_2 \rangle + 1)^d$. For $d=2$, the mapping is as:

$$\begin{aligned} \phi(x) = (&1, \sqrt{2}x_1, \cdots, \sqrt{2}x_p, x_1^2, \cdots, x_p^2, \\ &\sqrt{2}x_1x_2, \cdots, \sqrt{2}x_1x_p, \sqrt{2}x_2x_3, \cdots, \sqrt{2}x_{p-1}x_p) \end{aligned} \tag{14}$$

So the mode of polynomial SVM can be expressed as:

$$\hat{y}(x) = w_0 + \sqrt{2}\sum_{i=1}^{p} w_i^{(1)} x_i + \sum_{i=1}^{p} w_{i,i}^{(2)} x_i^2 + \sqrt{2}\sum_{i=1}^{p}\sum_{j=i+1}^{p} w_{i,j}^{(2)} x_i x_j \tag{15}$$

where the model parameter are:

$$w0 \in \mathbb{R}, \quad w^{(1)} \in \mathbb{R}^p, \quad w^{(2)} \in \mathbb{R}^{p \times p} \tag{16}$$

Compare the polynomial SVM with FM, we can find both model all nested interactions up to degree $d=2$. The main difference between polynomial SVM and FM is the parameterization. All interaction parameters are independent for SVM while the interaction parameters of FM are factorized and $<v_i, v_j>$ and $<v_i, v_l>$ depend on each other as they overlap and share parameter v_i.

High dimension of textual feature is main challenge for the stock market prediction based on social media. When the high dimension makes data linearly separated, prior researches show that linear kernel performs well [8]. Because linear kernel ignores the interactions between features, polynomial kernel is used model all nested interactions. However, so many independent parameters are difficult to estimate exactly for the polynomial kernel. In contrast, FM just models pairwise interaction and uses factorization parameterization to estimate parameters reliably. From the perspective of model formulation, FM is better than SVM with linear and polynomial kernel for stock market prediction.

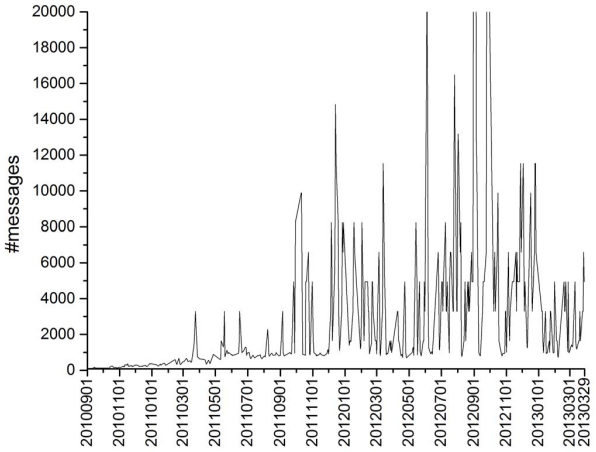

Fig. 1. Number of messages which include"上证指数(Shanghai Composite Index)"

IV. EXPERIMENT

A. Dataset and Evaluation Metric

We conduct our experiments on Sina Weibo, which is the most popular Chinese micro-blogging service and has more than 400 million registered users by the end of the third quarter in 2012. We download messages retrieved by the query "上证指数(Shanghai Composite Index)" on the search platform[3] of Sina Weibo from September 1st, 2010 to Match 29th, 2013. The daily amount of messages is shown in the Fig. 1. The amount of messages is small until September 2011. The good prediction of stock market is based on the enough amounts of messages about stock market in social media. To guarantee sufficient experimental data, we use data from September 29th, 2011 to Match 29th, 2013 which is through 361 trading days.

The statistics of the experimental data can be summarized in Table I. "#day", "#msg" and "#user" denote the number of trading days, messages and users respectively. "avg" means the average value. There are 256,691 messages collected in 361 trading days, and the average number of message per day is about 711. Of these messages, the average length of message is 36 words.

The number of unique users per day is shown in Fig. 2 and the average number of unique users per day is 520. We can find that there are many different users who discuss the stock market in social media per day. So many messages from different users can make the prediction more reliable.

[3] http://s.weibo.com

In contrast to financial news articles, social media is better information source. We summarize the advantages of social media as follows. Firstly, social media has become ubiquitous and important for social networking and content sharing for millions of users with different backgrounds. Secondly, social media has characteristics of the timely provision of new content and quick interaction among users. Thirdly, the intensive publishing and interaction can be viewed as a measure of users' attention towards a large range of news about stock market. Fourthly, the length limitation of posted message lets users write them efficiently and contains less noise data. Fifthly, the repost mechanism (e.g. called retweet in Twitter) allows users to share information with their followers and accelerates the spread of information in social media. The useful information is more likely to be reposted. Social media includes not only information of financial news but also discussions about the financial news, and provides richer and timely information for the stock market prediction.

TABLE I. STATISTICS OF EXPERIMENTAL DATA

#day	#msg	#user	avg #msg per day	avg #word per msg	avg #user per day
361	256,691	43,316	711	36	520

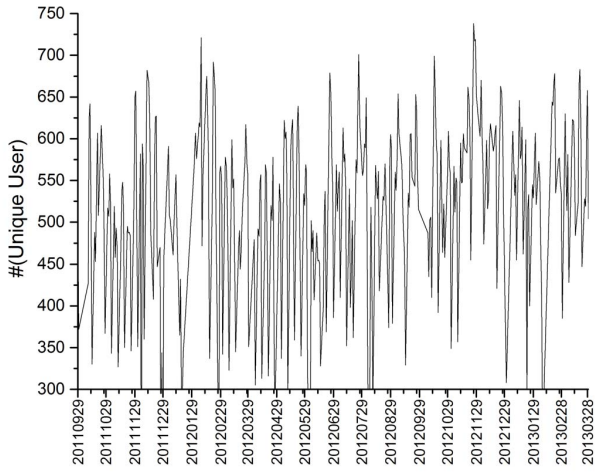

Fig. 2. The number of unique users per day

We collect the closing prices of Shanghai Composite Index from Yahoo Finance[4] during the period September 1st, 2010 to Match 29th, 2013. The closing price on each day is shown in Fig. 3.

Two evaluation metrics are used in our experiments. The first one is the accuracy. The accuracy of the prediction is obtained by checking whether the direction of the predicted price is the same as the actual trend. For instance, if the prediction for the upcoming trend is up and the upcoming trend is really rising, then we say that the prediction is correct; otherwise, if the prediction is down, but the upcoming trend is rising, then we say that the prediction is wrong. Here we use the following formula to measure the accuracy of prediction, which has adopted by many previous works[35].

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{17}$$

where *tp* and *tn* refer to the number of true positives and true negatives respectively. *fp* and *fn* is the number of false negatives.

While the accuracy just considers the direction rather than the magnitude, this leads us to evaluation measure using another evaluation method i.e. simulated trading system. The trading system follows simple trading rules and invests ¥1,000 per trade. The rules of our trading system are greedy to maximize daily trading profit. While our simulated trading system will buy the stock at the end of previous day if the predicted price will rise, nothing is done if the predicted price will drop. Any bought stocks are then sold at the end of current day. We also assume a zero transaction cost which is consistent with the prior studies [17, 24]. The final amount of earned money is our profit, which is used to another evaluation metrics. In our experiment, the significance ($p<0.05$) is tested using a 2-tailed paired t-test. For one evaluation metric, the significance test is used to compare whether a result is better than another significantly.

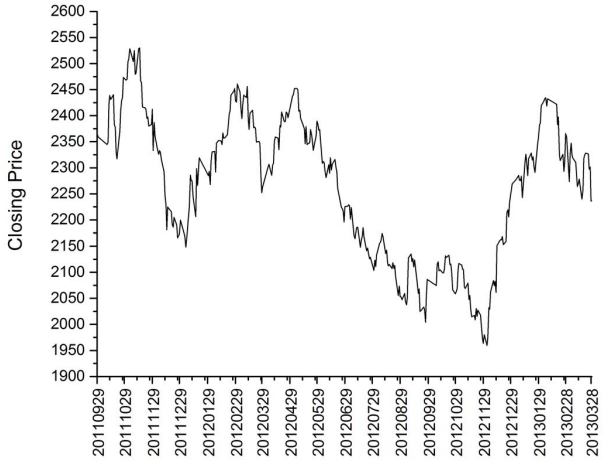

Fig. 3. Closing price of Shanghai Composite Index

B. Experimental Setup

In this study, our goal is to predict the closing price of Shanghai Composite Index. The period is from September 29th, 2011 to Match 29th, 2013, which is through 361 trading days. We used 1 or 0 to label the up or down of Shanghai Composite Index compared with the previous day. The data of previous 261 days is kept for training and remaining 100 days is used for testing.

Stanford word segmenter[5] is used to split Chinese text into a sequence of words. We remove standard stop words and punctuation marks, and filter out words that occur in less than 10 messages of training data. The bag of words is used as

[4] http://finance.yahoo.com/stock-center/

[5] http://nlp.stanford.edu/software/segmenter.shtml

features. Similar to the normalized document frequency, we calculate the normalized message frequency for each word as its weight in Eq.(18).

$$Weight(w,t) = \frac{mf(w,t)}{N(t)} \tag{18}$$

where *mf(w,t)* denotes the message frequency of word *w* on day *t*, and *N(t)* is the total number of message on day *t*. For example, there are 1000 messages in one day and 800 messages that include a word whose feature value is 0.8.

C. Model Performance

We compare our method against some baselines, which include the Simple Moving Average, Supervised Latent Dirichlet Allocation, Logistic Regression and SVM. All models' parameters are specified by the 5-fold cross validation on the training data.

1. Simple Moving Average (SMA): The moving average is an important and reliable statistical measurement, which can track the trend of stock market automatically based on historical price information. A common moving average model is the simple moving average defined as:

$$SMA(t) = \frac{\sum_{i=1}^{k} X_{t-i}}{k} \tag{19}$$

where t is the target date and k denotes the number of pervious days. X_{t-i} is the closing price of (t-i)-th day. SMA predicts the closing price on t-th day by using the past k days of data. We set $k = 3$ in our experiments.

2. Supervised Latent Dirichlet Allocation(sLDA): Proposed by [2], sLDA is a statistical model of labeled documents. Under this model, each document and its label arise from the two-stage generative process. In our experiment, all messages on one day are regarded as one document.

3. Logistic Regression (LR): Logistic regression model is a type of GLM, which can be used to binary classification, as shown in Eq.(9).

4. SVM: SVM is often used to stock market prediction due to its excellent generalization performance [18]. Linear kernel is used in our experiment.

5. Up Bound(UB): In order to evaluate our model comprehensively, we assume there is a perfect model which can achieve 100% accuracy, and the up bound can be obtain by this model. We compare the performance of these models with the up bound, and know to how much extent each model can perform well.

The simple moving average assumes that the future closing price is associated with prices of past k days regardless of the new information related with the stock market. Table II shows that SMA gets bad performance. There are similar results even if we increase the number of past days. The accuracy based on history prices is similar to the random guess, which reveals that historical prices have less impact on stock market in the future.

sLDA is significantly better than SMA. The prediction can benefit from analyzing topics related to the stock market, and users' discussion in social media can indicate the trend of stock market to some extent. We analyze why FM is better than LR and SVM from perspective of model formulation in Section 2. It is interesting to find that although LR, SVM and FM have similar accuracy, FM performs significantly better than LR and SVM in the metric of profit. Therefore, FM is more likely to accurately predict the larger increase or reduction of stock price than LR and SVM. In other words, FM has advantage in predicting the higher volatility in the stock market while it makes mistakes for the lower volatility. In contrast to the up bound, we find that FM has comparative good performance.

TABLE II. EVALUATION RESULTS OF DIFFERENT MODELS

	SMA	sLDA	LR	SVM	FM	UB
Accu	49%	66%	76%	77%	81%	100%
Profit	75.55	208.3	339.2	390.4	405.4	495.9

We find that SVM with polynomial kernel does not work for stock market prediction. The high dimension of features makes p a large number, and the 2-degree polynomial kernel maps the feature into a higher dimensional space. It is difficult to estimate so huge number of independent parameters reliably by the relatively small training data.

D. Textual Representation

Following the work [35], we examine the performance of different models for stock market prediction using several different textual representations: bag of words, noun phrases, verb phrases and named entities. Stanford part-of-speech tagger[6] is used to assign parts of speech to each Chinese word, such as noun, verb and etc. Stanford named entity recognizer[7] is used to label Chinese named entities in messages and the predetermined categories include location, organization, person and etc.

As shown in Table III, different textual representations have greater impact on logistics regression model, while SVM and FM are less affected by the textual representations.

We find that the textual representation of named entity fails for stock market prediction because most users, who discuss stock market in social media, neither make use of named entities nor repost messages that contain named entities. Even if they use named entities when posting messages, these named entities are not associated with the stock market. It is reasonable to believe that the composite index prediction differs from that of company because the later can benefit from the recognition of company name.

E. Sensitivity to Dimensionality

The factorization dimensionality k is an important parameter. When applying the FM to stock market prediction, we assume that feature interactions play a significant role to

[6] http://nlp.stanford.edu/software/tagger.shtml
[7] http://nlp.stanford.edu/software/CRF-NER.shtml

improve prediction. To validate this assumption, we train FM model with different *k*. The bag-of-words features are used as textual representation.

As shown in Fig. 4, when the dimensionality changes the accuracy is almost similar while the profit varies in a wide range. We observe that FM gets the most profit with *k=20*. In addition, the improvements over FM with *k=0* are significant, which shows that 2-degree FM is more effective in stock market prediction because interactions between features are taken into account. On the other hand, when *k* becomes large, the performance of 2-degree FM drops while it is still better than the FM with *k=0*. Therefore, it is important to select an appropriate *k*. The factorization parameterization also makes 2-degree FM more adaptive by adjusting the number of factors *k* for specific applications.

TABLE III. RESULT OF DIFFERENT TEXTUAL REPRESENTATION

	LR	SVM	FM
bag of words	339.26(76%)	390.41(77%)	402.94(80%)
noun phrases	342.87(75%)	398.89(79%)	394.68(78%)
verb phrases	358.39(75%)	393.74(79%)	396.84(79%)
noun+verb	339.93(75%)	394.28(78%)	387.29(78%)

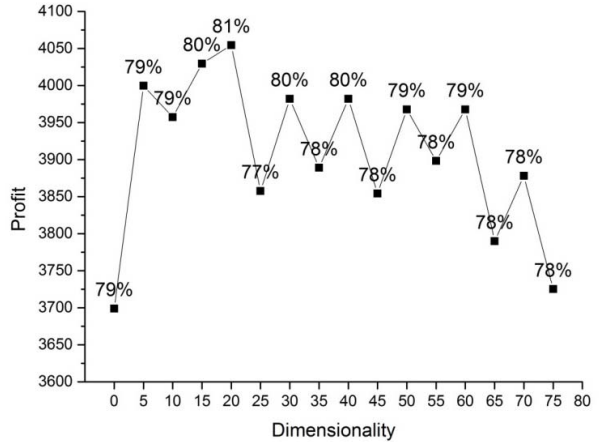

Fig. 4. Profit with different dimensionality. The percentage above points denotes the accuracy.

V. CONCLUSION

We exploit social media for stock market prediction with factorization machine that can be used to produce more accurate prediction. From the perspective of model formulation, we derive factorization machine's relationship with GLM and SVM and believe factorization machine may be more accurate than others. Our experiments on real-world data demonstrate that factorization machine can significantly improve the prediction of stock market against several not-trivial models. In addition, we investigate how textual representation and the dimensionality influence the prediction, and find that factorization machines are stable. Although our work has been done in the context of Sina Weibo, we expect the same results would hold for many other similar social media such as Twitter and Facebook.

For future work, we will find more microblogs in social media that are associated with stock market by increasing search keywords. We just use Factorization Machines in a new application, while the adaptation of factorization machines to improve stock market prediction is another direction. Our method requires that each day has sufficient number of messages to model stock market. How to model stock market with small data will be another direction of our future work.

VI. ACKNOWLEDGMENTS

Research was partially supported by National Natural Science Foundation of China (No. 61170184), the Tianjin Municipal Science and Technology Commission (No. 13ZCZDGX02200, 13ZCZDGX01098), the Research Foundation of Ministry of Education-China Mobile (No. MCM20130381), and Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Asur, S., Huberman, B.A. Predicting the future with social media. In Proceeding of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 492-499, 2010.

[2] Blei, D.M.A.M. Supervised topic models. In Proceedings of NIPS, 121-128, 2007.

[3] Bollen, J., Mao, H., Zeng, X. Twitter mood predicts the stock market. Journal of Computational Science, 2 (1): 1-8, 2011.

[4] Bothos, E., Apostolou, D., Mentzas, G. Using Social Media to Predict Future Events with Agent-Based Markets. Intelligent Systems, IEEE, 25 (6): 50-58, 2010.

[5] Chang, P., Wang, D., Zhou, C. A novel model by evolving partially connected neural network for stock price trend forecasting. Expert Systems with Applications, 39 (1): 611-620, 2012.

[6] Choi, H., Varian, H. Predicting the present with google trends. Economic Record, 88 (s1): 2-9, 2012.

[7] Dai, W., Wu, J., Lu, C. Combining nonlinear independent component analysis and neural network for the prediction of Asian stock market indexes. Expert Systems with Applications, 39 (4): 4444-4452, 2012.

[8] Dumais, S., Chen, H. Hierarchical classification of Web content. In Proceedings of SIGIR, 256-263, 2000.

[9] Fama, E.F. The behavior of stock-market prices. The journal of Business, 38 (1): 34-105, 1965.

[10] Fung, G.P.C., Yu, J.X., Lu, H. The Predicting Power of Textual Information on Financial Markets. IEEE Intelligent Informatics Bulletin, 5 (1): 1-10, 2005.

[11] Gallagher, L.A., Taylor, M.P. Permanent and temporary components of stock prices: Evidence from assessing macroeconomic shocks. Southern Economic Journal: 345-362, 2002.

[12] Gilbert, E., Karahalios, K. Widespread Worry and the Stock Market. In Proceeding of ICWSM, 59-65, 2010.

[13] Gruhl, D., Guha, R., Kumar, R., Novak, J., Tomkins, A. The predictive power of online chatter. In Proceedings of SIGKDD, 78-87, 2005.

[14] Hong, L., Doumith, A.S., Davison, B.D. Co-factorization machines: modeling user interests and predicting individual decisions in Twitter. In Proceedings of WSDM, 557-566, 2013.

[15] Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A. Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology, 60 (11): 2169-2188, 2009.

[16] Koren, Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In Proceedings of SIGKDD, 426-434, 2008.

[17] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., Allan, J. Language models for financial news recommendation. In Proceedings of CIKM, 389-396, 2000.

[18] Lee, M. Using support vector machine with a hybrid feature selection method to the stock trend prediction. Expert Systems with Applications, 36 (8): 10896-10904, 2009.

[19] Li, X., Wang, C., Dong, J., Wang, F., Deng, X., Zhu, S. Improving stock market prediction by integrating both market news and stock prices. In Proceedings of DEXA, 279-293, 2011.

[20] Liu, Y., Huang, X., An, A., Yu, X. ARSA: a sentiment-aware model for predicting sales performance using blogs. In Proceedings of SIGIR, 607-614, 2007.

[21] Majhi, R., Panda, G., Sahoo, G. Development and performance evaluation of FLANN based model for forecasting of stock markets. Expert Systems with Applications, 36 (3): 6800-6808, 2009.

[22] Malkiel, B.G. A Random Walk Down Wall Street. W.W. Norton & Company, New York, 1973.

[23] Mishne, G., de Rijke, M. A study of blog search. In Proceedings of the 28th European conference on Advances in Information Retrieval, 289-301, 2006.

[24] Mittermayer, M. Forecasting intraday stock price trends with text mining techniques. In Proceedings of HICSS, 10, 2004.

[25] Nofsinger, J.R. Social mood and financial economics. The Journal of Behavioral Finance, 6 (3): 144-160, 2005.

[26] O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In Proceedings of ICWSM, 122-129, 2010.

[27] Oh, C., Sheng, O. Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement. In Proceedings of ICIS, 2011.

[28] Qian, B., Rasheed, K. Stock market prediction with multiple classifiers. Applied Intelligence, 26 (1): 25-33, 2007.

[29] Qiang, R., Liang, F., Yang, J. Exploiting ranking factorization machines for microblog retrieval. In Proceedings of CIKM, 1783-1788, 2013.

[30] Rendle, S. Factorization machines. In Proceedings of ICDM, 995-1000, 2010.

[31] Rendle, S. Factorization machines with libFM. ACM Transactions on Intelligent Systems and Technology, 3 (3): 57, 2012.

[32] Rendle, S., Gantner, Z., Freudenthaler, C., Schmidt-Thieme, L. Fast context-aware recommendations with factorization machines. In Proceedings of SIGIR, 635-644, 2011.

[33] Rendle, S., Schmidt-Thieme, L. Pairwise interaction tensor factorization for personalized tag recommendation. In Proceedings of WSDM, 81-90, 2010.

[34] Rout, M., Majhi, B., Mohapatra, U.M., Mahapatra, R. Stock indices prediction using radial basis function neural network. in Swarm, Evolutionary, and Memetic Computing, Springer, 2012, 285-293.

[35] Schumaker, R.P., Chen, H. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems, 27 (2): 12, 2009.

[36] Shen, W., Guo, X., Wu, C., Wu, D. Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm. Knowledge-Based Systems, 24 (3): 378-385, 2011.

[37] Srebro, N., Jaakkola, T., Others. Weighted low-rank approximations. In Proceeding of ICML, 720-727, 2003.

[38] Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proceedings of ICWSM, 10: 178-185, 2010.

[39] Wang, F., Liu, L., Dou, C. Stock Market Volatility Prediction: A Service-Oriented Multi-kernel Learning Approach. In Proceedings of SCC, 49-56, 2012.