2014 IEEE 11th International Conference on e-Business Engineering

Public Sentiment Analysis in Twitter Data for Prediction of A Company's Stock Price Movements

LI Bing[1] Keith C.C. Chan[1] Carol OU[2]

[1]Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon,
Hong Kong
{csbingli, cskcchan}@comp.polyu.edu.hk

[2] Department of Management
Tilburg School of Economics and Management
Tilburg University
Tilburg, Netherlands
carol.ou@uvt.nl

***Abstract*—There has recently been some effort to mine social media for public sentiment analysis. Studies have suggested that public emotions shown through Tweeter may well be correlated with the Dow Jones Industrial Average. However, can public sentiment be analyzed to predict the movements of the stock price of a particular company? If so, is it possible for the stock price of one company to be more predictable than that of another company? Is there a particular kind of companies whose stock price are more predictable based on analyzing public sentiments as reflected in Twitter data? In this article, we propose a method to mine Twitter data for answers to these questions. Specifically, we propose to use a data mining algorithm to determine if the price of a selection of 30 companies listed in NASDAQ and the New York Stock Exchange can actually be predicted by the given 15 million records of tweets (i.e., Twitter messages). We do so by extracting ambiguous textual tweet data through NLP techniques to define public sentiment, then make use of a data mining technique to discover patterns between public sentiment and real stock price movements. With the proposed algorithm, we manage to discover that it is possible for the stock price of some companies to be predicted with an average accuracy as high as 76.12%. In this paper, we describe the data mining algorithm that we use and discuss the key findings in relation to the questions posed.**

Keywords—social media; Twitter; stock market; data mining

I. INTRODUCTION

Social media can be defined as the media that consists of a network of relationships, in which the nodes are the actors and the edges are the relationships between those actors [1]. The analysis of social media can be traced back to the time before the invention of World Wide Web. More recently famous researchers like Milgram [2] implemented their analysis and mining techniques on social media in 20th century. However, as a result of the modern IT techniques that developed rapidly, social media has exploded at an exponential rate. Functioning as the online forums, social media such as Twitter, Flickr, LinkedIn and Facebook enables users to post content, bookmark them, share them, review them and connect at a rapid rate. More specifically, some social media sites like Facebook and Twitter are explicitly designed for social interactions, while others like Flickr are more tendentiously designed for content sharing. Attributed to social media's high level of ease to use, reach, richness and immediacy, public opinions and discourses are changing rapidly, and its influences are extended to various domains such as politics, environment, entertainment industry, stock market, etc. The availability of massive amounts of data has drawn great attentions on researching social media statistically or in a more general sense, scientifically and meaningfully. In another words, the new era of analyzing and mining the data from social media is data centric. For instance, as a typical example of social media, Twitter is a micro-blogging application that allows interested users to follow and comment other users' thoughts or some events in their lives, in real time [3]. As one of the most popular social media, more than millions users post over 140 million tweets every day [4]. This situation makes Twitter like a corpus with valuable data, and has attracted much attention from a various fields of researchers in recent years.

Motivated by the large amount of valuable knowledge existed in social media, in this article, we attempt to use the Twitter data as our information corpus to predict the share prices of specific stocks in American stock markets. Stock market price prediction has always been a fancy issue over past two decades, but with poor practical solutions [5]. For instance, some of their works, based on the random walk theory, have only achieved poor predict accuracy. Others focused on the detailed financial news about the listed companies [6], like the classic Efficient Market Hypothesis (EMH) [7]. Commonly speaking, EMH in analyzing stock market prices are mainly based on the financial news. We argue that the news is actually hard to predict, and therefore it is unreasonable to predict something based on unpredictable factors. Furthermore, even some works like [8] and [9] have claimed a high predicted accuracy, those works are lack of generalization because too many specified conditions were required to make the predictions. It has been widely accepted by economic specialists that there is a potential connection between the company's stock price and the published information about it [10]. Since the data in social media can be collected by technical methods, we can analyze the public views about the products of listed

978-1-4799-6563-2/14 $31.00 © 2014 IEEE
DOI 10.1109/ICEBE.2014.47

companies, which can provide us a novel opportunity to predict the share price in stock market.

Followed by the above motivation, we propose to find a solution for mining the data from Twitter database in order to link these ambiguous data with the trend of American stock market price by applying the sentiment analysis and data mining algorithms. Firstly, we consider each Tweets structure as the combination of several words and phrases, and apply the NLP techniques to classify the Tweets sentiment into five categories (Positive$^+$, Positive, Neutral, Negative, Negative$^-$). Then we apply chi-square test and adjusted residual to identify the interested patterns between public sentiments and stock market prices. Through the experimental part, we select 30 representative listed companies from different industries as the target and retrieve almost 15 million records of Twitter data which may mention about these companies, for instance, their products. For different companies, we use different strategy to retrieve the related Tweets, like the company "Apple Inc" which listed code is "AAPL", simply to use its code as the keyword to collect its related Tweets is obvious unreasonable, but to use "iPad", "iTunes" and "iPhone", and their characteristics such as "CPU Speed", "Color", etc.. We apply the proposed algorithm to calculate the Twitter sentiment degree of these listed companies and collect their real stock price values in NASDAQ Stock Market and New York Stock Exchange from Yahoo! Finance, in order to find the correlation between Twitter public mood and the trend of real stock market price. According to the experimental result, we find that through our proposed algorithm, the real stock price can be predicted with the 81.45% predict accuracy.

The rest of this article is organized as follows. Section 2 introduces the background of the proposed work and provides a survey of related literature to the identified problem. Section 3 states and explains our proposed methods to tackle the research problem. Then, section 4 present the experimental result and discussion to show that the real stock price could be predicted and our proposed performed well with the high predict accuracy. Lastly, section 5 concludes this article.

II. Related Works

Although social media has been increasingly a research topic, not many studies focus on data-timeline based nature of social media and the ambiguous information embedded in social media. Jansen and his colleagues [11] examined social media as the platform of word-of-mouth advertisements, and considered particular brands and products when investigating the structure of those postings at social media and the changes in sentiments. This research provides insights on how to analyze ambiguous information at social media. However the predictive aspect of social media like Twitter remains unexplored. Moreover, the nature of social media data is based on time line, ambiguous and enormous. However, most research on social media has yet to properly handle this nature of social media.

There are some prior works on analyzing the correlation between blogs and reviews' mentions and predictable performance. For instance, Gruhl and his colleagues [12] have shown how to generate automated queries for mining data from blogs, in order to predict spikes in the sales volume of books. Meanwhile, although there are considerable research on predicting the sales volume of movie sales using the data from social media, all of them have used the meta-data information about the movies themselves to perform the forecasting, including MPAA rating, movies genre, the number of screens on which the movie debuted, running time, released date and the presence of particular actors or actresses in the cast. Others like Joshi and his colleagues [13] used linear regression from textual data and metadata features to predict earnings about the posted movies. Sharda and Delen [14] have treated the prediction problem as a traditional classification problem and applied the neural networks to classify movies into categories ranging from the 'blockbuster' to the 'flop'. Apart from the fact that they predicted the ranges instead of the actual numbers, the best accuracy of their models was fairly low.

Focusing on mining temporal data from social media, Asur and Huberman [15] from HP Labs proposed a model, based on the popular rate at which tweets are created about particular topics, to predict the actual box office revenue of a movie, but the prediction was conducted only after the movie is released several weeks after in order to collect enough opinions. The dataset which they used was obtained by crawling hourly feeding data from Twitter with Twitter Search Api. This ensured the correct timestamp, author and tweet text for their analyses. In order to ensure the obtained tweets that all referring to a movie, they used keywords present in the movie title as the search arguments. This yielded to 2.89 million tweets which referred to 24 different movies released over a period of three months. The authors also claimed that the predictions from tweets were more accurate than any other methods of forecasting like Hollywood Stock Exchange index [16] and News-based Prediction [17]. Although Asur and. Huberman' treated the data from social media as the temporal data mining issue, their method of text analysis was relatively simple, as they overlooked the ambiguous information in social media.

Focusing on the stock market prediction, Bollen and his colleagues have proposed two mood tracking tools, namely OpinionFinder and Google-Profile of Mood States, to analyze the tweets' mood, with 6 dimensions covering calm, alert, sure, vital, kind, and happy [9]. They made used of these two tools to measure variations in the public moods from tweets which were submitted to the Twitter database from February 28, 2008 to December 19, 2008. More specifically, the first tool, OpinionFinder, could analyze the textual content of tweets which were submitted on a given day in order to classify them into positive vs. negative public moods based on the daily time series. The second tool, GPOMS, could similarly analyze the textual content of tweets in order to generate the six-dimensional public mood based on the daily time series in order to provide a more detailed view of the changes along a variety of different mood dimensions. The authors also applied the Granger Causality Analysis [18] and Self-Organizing Fuzzy Neural Network to examine the hypothesis that public mood states, as measured by the time series of OpinionFinder and GPOMS moods, are predictive of up or down changes in DJIA closing value. Their analysis achieved an accuracy rate of 87.6% and a reduction of the Mean Average Percentage Error by more than 6%. Followed by their research, Mittal and Goel [19] made a similar research on predicting Dow Jones Industrial Average trend through tweets, but mainly focused on the financial news related tweets, with a prediction accuracy rate of 87%, as same as Bollen et al's. However, there are at least two good reasons to further extend their research based on the following reasons. Firstly, the datasets of their research was too small and the time period which they selected was too short. Plus, they all ignored the geography issue. It is unreasonable that people in Asia could influence with the same weight on Dow Jones Industrial Average with people in U.S. In order to address these two issues, we will show how our propose algorithms and architecture can deal with a big dataset and apply more reasonable analyses when compared to their methods in the next section.

III. Proposed Method

Through evaluating the related works, we list the following research problems to be solved in this section:

1) The text data embedded in social media is commonly unstructured and ambiguous and thus, generally speaking, the data for one attribute will not be clear enough and sufficient to provide appropriate information for analysis. The data structure in social media is multilayered, which means that the attributes of the dataset are always connected with each other in a hierarchical structure. For instance, "Apple Inc.", "iPhone 5s", "MacBook Pro", "Battery", "Color" can be used as keywords and can all be treated as attributes for all collected social media data. However, the real situation is that "iPhone 5s" and "MacBook Pro" are products manufactured by the company "Apple Inc.", and furthermore, "Battery" and "Processing Speed" are features describing the products "iPhone 5s" and "MacBook Pro". The previous social media mining research have not respect the fact by just roughly defining the relationship between attributes in the whole social media dataset. Hence, a method to define and fully identify this embedded information, and the multilayered attributes, in the social media environment is important for use in social media data mining.

2) The attributes mentioned in social media data may be connected with each other and may reflect hierarchical structures. More formalism, in real situations some specific features may have internal connections with the selected target. But these connections may also not be evident, especially in the data format used in the social media environment. Therefore, an algorithm to measure such connections and to mine how these specific features affect the target significantly is important for problems of both association mining and classification mining. For instance, the algorithm should be able to identify qualified patterns of association between all attributes at all layers to fully represent the association problems. These can include how significantly the lower layered attributes have affected the top layered attributes. Moreover, if a classification problem arises in the mining of opinions about a top layered attribute, such as "Apple Inc.", the algorithm should be able to accurately measure all of the weights of the linked lower layered attributes. It can then summarize this supporting information and use it to classify the opinions about the top layered attribute into several categories.

In order to solve the above mentioned research problems, this section will introduce our proposed algorithm to mine the multi-layered social media data in details. It mainly consists of two steps, first, it applies NLP technique to extract the sentiment word list to classify each tweets into sentiment categories. Then, it applies data mining technique to discover interested association rules between each defined attributes, thus to predict the stock price movements by public sentiment.

A. Classify Tweets into Sentiment Categories

As each origin Tweets is one or several sentences which can be considered like the format "A+B+C+D+E+B…+R", "A", "B", "C"… "R" means a different word, the basic method to classify these Tweets is firstly create the word list to compute each word's sentiment value, then, according to each word's value and consider in potential phrases and expressions, to classify the Tweets sentiment categories. In order to fit the hot topics through the social networks to construct the sufficient sentiment word list, we parallel select 10 times of training samples. For each selection, we firmly select 5000 words as the total numbers of training samples:

$$a = f(x, S = 5000) = \begin{cases} \frac{S}{\sum_i^n T_i}, & \sum_i^n T_i > S \\ \sum_i^n T_i, & \sum_i^n T_i \leq S \end{cases} \tag{1}$$

$$S' = a \times \sum_i^n T_i\,,\ S' \geq S \tag{2}$$

For those selected words, Vector Space Model, which is an algebraic model to represent text information as vectors of identifiers, and TF-IDF(term frequency-inverse document frequency) model [22], which is a classic VSM model are used to formulate their weight in the real linguistic environment, as the following formula shows.

$$W(i, \vec{d}) = \frac{tf(i,\vec{d}) \times \log(\frac{N}{n_i} + 0.01)}{\sqrt{\sum_{i \in \vec{d}} [tf(i,\vec{d}) \times \log(\frac{N}{n_i} + 0.01)}} \tag{3}$$

In the above formula, the $W(i, \vec{d})$ means the weight of the word i in Tweet $\vec{d}$, $tf(i, \vec{d})$ means the frequency of the word i in Tweet $\vec{d}$, and N mean the total numbers of training samples. According to the value of $W(i, \vec{d})$, we have firmly selected 8892 eligible sentiment words which consists of 803 Positive$^+$ words, 3572 Positive words, 2152 Neutral words, 1476 Negative words and 979 Negative$^-$ words, to create the word list. We also used the SentiWordNet 3.0, which is a lexical application for sentiment classification, to calculate the specific words' sentiment value from the created word list. In SentiWordNet 3.0, each word has been given three kinds of notions: positivity, negativity and neutrality. Cross our system, by matching inputted word with SentiWordNet 3.0, it will give each numerical score of the word, each score ranges from 0.0 to 1.0, and their sum is 1, and save the results in our local server, as the following table 1 shows:

TABLE 1 EXAMPLES OF SENTIMENT WORDS

P	O	N	Text
0.5	0.25	0.25	hopeful#1
0	0.65	0.375	promising#2 hopeful#2 bright#10
0	1	0	sympathetic#1
0.375	0.375	0.25	sympathetic#2
0.625	0.375	0	sympathetic#3 openhearted#1 large-hearted#1 kindly#1 good-hearted#1

According to the specific words' sentiment value from SentiWordNet 3.0, we use the following formula (4) to classify each tweets into five categories. In the formula, the $(\mathrm{P_i} - \mathrm{N_i})$ means the polarity value of the sentiment word in a particular tweets, i means the rank of the target word in the set of its synonyms. The rank is more forward, the weight of its polarity value is higher, so we use $\frac{1}{\mathrm{i}}$ to measure it.

$$\mathrm{W_{score}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{(\mathrm{P_i} - \mathrm{N_i})}{\mathrm{i}}}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{1}{\mathrm{i}}} \tag{4}$$

$$\mathrm{W_{tag}} = \left\{ \begin{array}{ll} \text{Positive}^{+} & \mathrm{W_{score}} \geq 0.75 \\ \text{Positive} & 0.25 \leq \mathrm{W_{score}} < 0.75 \\ \text{Neutral} & -0.25 < \mathrm{W_{score}} < 0.25 \\ \text{Negative} & -0.75 < \mathrm{W_{score}} \leq -0.25 \\ \text{Negative}^{-} & \mathrm{W_{score}} \leq -0.75 \end{array} \right\}$$

Through the calculated sentiment classification of each frequent word, we further use C-SVC to classify each tweets into five pre-defined sentiment categories. More specifically, "one-against-one" approach [24] is used to solve the M-class classification problem. We represent each tweets by the word from the created word list and its sentiment values as the following training vectors to train the input tweets, in order to assign them with sentiment possibilities according to 5 defined sentiment categories:

$$word_i value_j\colon W_{i,j} \cdots word_n value_M\colon W_{n,M}\ C_k$$
$$where\ i = 1, \ldots, n; j = 1, \ldots, M; k = 1, \ldots, M; value_j,$$
$$C_j \in \{Negative^{-}, Negative, Neutral, Positive, Positive^{+}\}$$

B. Association Rule Mining

Through the last step, each tweets has been classified into five sentiment categories, the next step is to discover patterns between the public sentiment and the stock price movements. We first give the following definitions to formalize this classification problem. Suppose that there is a dataset LD constructed by a set of attributes $\mathrm{E} = \{E_1, E_2, \cdots, E_n\}$, where E_j, $\mathrm{i} = 1, 2, \ldots, \mathrm{n}$ can be quantitative or categorical; and a set of ordered sequence objects $\mathrm{S} = \{S_1, S_2, \cdots, S_m\}$, in which $1 \leq \mathrm{i} \leq \mathrm{n}$ and $1 \leq \mathrm{j} \leq \mathrm{m}$. For any records, $\mathrm{d} \in \mathrm{LD}$, $\mathrm{d}[E_j]$ denotes the value for the attribute E_j. Furthermore, it is supposed that each element in sets E and S is independent, $E_i \cap E_j = \emptyset$ and $S_i \cap S_j = \emptyset, i \neq j$. Hence, we can consider that each attribute E_j represents the multi-layered features of target object, each attribute S_j represents the time slot of the tweets published date, as the following figure 1 shows, for instance, in order to analyze the listed company "Apple.Inc", each defined attributes could be the products of this company and their characteristics.

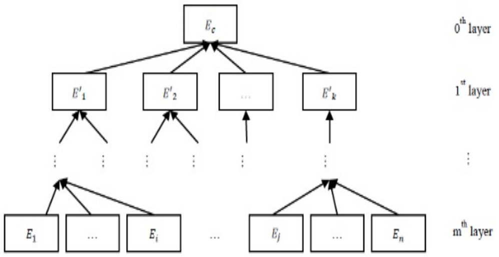

Fig. 1. Conceptual Structure of Multi-layerd Attributes to Represent the Object

As hypothesized above, any of the measured attributes in attribute set E is independent with each other, this means that the null hypothesis is proved, and hence the chi-square test is proposed to evaluate the relationships between attributes and be used for further measuring the weight of the revealed association rules. If the problem is to determine whether two attributes are associated, such as E_j and E_k as described above, we merge all the tweets using the set of ordered sequence objects S, to summarize all the information related to each attribute. Then the degrees of grouped sentiment possibilities for these two attributes are used to construct chi-square test table as shown below. More specifically, two sets, $\{\mathrm{j1, j2}, \cdots, \mathrm{jr}\}$ and $\{\mathrm{k1, k2}, \cdots, \mathrm{kr}\}$, are defined to represent the possibilities of each sentiment categories calculated from the first step, for E_j and E_k, as figure 2 shows.

In the chi-square test table, $\mathrm{f}(j_{i,1}, k_{i,1})$ measures the minimum value of $j_{i,1}$ and $k_{i,1}$, and $\sum_{i=1}^{m} \mathrm{f}(j_{i,1}, k_{i,1})$ measures all the information for the set of ordered sequence objects S. Following the pre-defined attributes and values, the chi-square statistic can be further defined as formula (5) shows.

$$\begin{array}{rc} & \begin{array}{cccc} L_{j1} & L_{j2} & \cdots & L_{jr} \end{array} \\ \begin{array}{r} S_1\{ \\ \textit{tweets of } E_j \Rightarrow S_2\{ \\ \vdots \\ S_m\{ \end{array} & \left| \begin{array}{cccc} \overset{\frown}{j_{11}} & \overset{\frown}{j_{12}} & \cdots & \overset{\frown}{j_{1r}} \\ j_{21} & j_{22} & \cdots & j_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ j_{m1} & j_{m2} & \cdots & j_{mr} \end{array} \right| \end{array} \quad \begin{array}{rc} & \begin{array}{cccc} L_{k1} & L_{k2} & \cdots & L_{kr} \end{array} \\ \begin{array}{r} S_1\{ \\ \textit{tweets of } E_k \Rightarrow S_2\{ \\ \vdots \\ S_m\{ \end{array} & \left| \begin{array}{cccc} \overset{\frown}{k_{11}} & \overset{\frown}{k_{12}} & \cdots & \overset{\frown}{k_{1r}} \\ k_{21} & k_{22} & \cdots & k_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ k_{m1} & k_{m2} & \cdots & k_{mr} \end{array} \right| \end{array}$$

$$\Downarrow$$

$$\begin{bmatrix} \sum_{i=1}^{m} f(j_{i,1}, k_{i,1}) & \sum_{i=1}^{m} f(j_{i,1}, k_{i,2}) & \cdots & \sum_{i=1}^{m} f(j_{i,1}, k_{i,r}) & \sum_{d=1}^{r}\sum_{i=1}^{m} f(j_{i,1}, k_{i,d}) \\ \sum_{i=1}^{m} f(j_{i,2}, k_{i,1}) & \sum_{i=1}^{m} f(j_{i,2}, k_{i,2}) & \cdots & \sum_{i=1}^{m} f(j_{i,2}, k_{i,r}) & \sum_{d=1}^{r}\sum_{i=1}^{m} f(j_{i,2}, k_{i,d}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^{m} f(j_{i,r}, k_{i,1}) & \sum_{i=1}^{m} f(j_{i,r}, k_{i,2}) & \cdots & \sum_{i=1}^{m} f(j_{i,r}, k_{i,r}) & \sum_{d=1}^{r}\sum_{i=1}^{m} f(j_{i,r}, k_{i,d}) \\ \sum_{i=1}^{m}\sum_{e=1}^{r} f(j_{i,e}, k_{i,1}) & \sum_{i=1}^{m}\sum_{e=1}^{r} f(j_{i,e}, k_{i,2}) & \cdots & \sum_{i=1}^{m}\sum_{e=1}^{r} f(j_{i,e}, k_{i,r}) & \sum_{d=1}^{r}\sum_{i=1}^{m}\sum_{e=1}^{r} f(j_{i,d}, k_{i,e}) \end{bmatrix} = \left[t_{i,j}\right]_{(r+1)\times(r+1)}$$

Fig. 2. Constructing the Chi-square Test Table

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{r}\frac{\left(t_{i,j} - \left(t_{i,r+1} \times t_{r+1,j}\right) / t_{r+1,r+1}\right)^2}{\left(t_{i,r+1} \times t_{r+1,j}\right) / t_{r+1,r+1}} = \sum_{l=1}^{r}\sum_{k=1}^{r}\frac{t_{i,j}{}^2}{\left(t_{i,r+1} \times t_{r+1,j}\right) / t_{r+1,r+1}} - t_{r+1,r+1} \tag{5}$$

Using above formula, if the calculation result is $X^2 > X^2{}_{0.05}((r-1)(r-1))$, then it can be said that under the degree of freedom $\mathrm{d} = (\mathrm{r}-1) \times (\mathrm{r}-1)$ and the significance level $\alpha = 0.05$, the two attributes E_j and E_k are associated with each other; conversely, it can be said that under the degree of freedom $\mathrm{d} = (\mathrm{r}-1) \times (\mathrm{r}-1)$ and the significance level $\alpha = 0.05$, the two attributes E_j and E_k are independent from each other. Through this process, it can be identified whether two single attributes are associated or not. However, there is no embedded information about whether specific sentiment categories are associated exclusively with a particular attribute. For instance, if the public opinion about "Iphone5s_battery" is negative, does this imply that the public opinion about "Apple.inc" is also negative? To deal with this problem, adjusted residual [20] is introduced to measure the differences between attributes such as E_j and E_k with specific sentiment categories $L_x \in \{L_{j_1}, L_{j_2}, \cdots, L_{j_r}\}$ and $L_y \in \{L_{k_1}, L_{k_2}, \cdots, L_{k_r}\}$:

$$d_{xy} = \frac{Z_{xy}}{\sqrt{V_{xy}}} \tag{6}$$

Where V_{xy} is the maximum likelihood estimate of the variance of Z_{xy}:

$$V_{xy} = \left(1 - \frac{t_{x,r+1}}{t_{r+1,r+1}}\right)\left(1 - \frac{t_{r+1,y}}{t_{r+1,r+1}}\right) \tag{7}$$

$$Z_{xy} = \frac{t_{x,y} - \left(t_{x,r+1} \times t_{r+1,y}\right) / t_{r+1,r+1}}{\sqrt{\left(t_{x,r+1} \times t_{r+1,y}\right) / t_{r+1,r+1}}} \tag{8}$$

Based on the results of the last step, which has found association rules between attributes, those attributes that are not associated are ignored, for example as in the cases of E_j and E_k, and Formulae (6, 7, 8) are defined to further identify whether they are associated with specific linguistic terms. Corresponding with the pre-defined table for the chi-square test, each element from it can be explored using the matrix T to calculate the value of Z_{xy}, V_{xy}, and d_{xy}. For instance, if the calculation result is $\left|d(L_x,\ L_y)\right| > 1.96$ (the 95 percentiles of the normal distribution), it can be said that the discrepancy between $\Pr(L_x|L_y)$ and $\Pr(L_x)$ is significantly different and hence the association between L_x and L_y is considered interesting. More specifically, if $\mathrm{d}(L_x,\ L_y) > 1.96$, the presence of L_x implies the presence of L_y and it can be said that L_{j_m} is positively associated with L_{j_n}. If $\mathrm{d}\left(L_x,\ L_y\right) < -1.96$, and the absence of L_y implies the presence of L_x and it can be said that L_x is negatively associated with L_y. For instance, as in the previous example, the result can be calculated that if the public's opinion about "Iphone5s_battery" is positive$^+$ this is associated with the public's opinion about "Apple.inc" that would also be positive. Or, if the public's opinion about "Iphone5s_color" and "MacBook Pro_Battery" are both positive, this is associated with the public's opinion about "Apple.inc" that would also be positive$^+$, as shown in the association rules below:

$$IF \text{ Iphone5s\_battery } is\ Positive^+\ THEN \text{ Apple. inc is } Positive$$

$$IF \text{ Iphone5}s\_color\ is\ Positive^+\ THEN \text{ MacBook Pro\_Battery is } Positive$$

C. Classify Object into Sentiment Categories

Moreover, as described earlier, the data in the social media may have multilevel meanings. For example, if the problem is to identify the overall sentiment about the company "Apple Inc.", the calculated sentiment about the products of that company, such as the "iPhone" and "MacBook Pro", play a significant role and do have some weight. Also, detailed features of the products, such as "CPU performance" and "appearance" have an influence on the sentiment about the top level attribute "Apple Inc." As was discussed before, the method proposed there can increase the dimensionality of the social media data for the target object using both direct and indirect division operations to measure the relations between the multilayered attributes. Hence, using the identification of any association rules between attributes, the classification problem can be formalized as a multi-dimensional patterns classification problem. This can be dealt with by classifying the upper layer attributes into categories, according to the defined linguistic terms, and measuring the weights of all of the relevant rules related to the lower layer attributes. For this the association rule is $L_q \rightarrow C_q$, which $L_q = \{L_{q1},\ L_{q2},\ \cdots,\ L_{qm}\}$ is the sentiment set with defined sentiment categories $[Negative^-, Negative, Neutral, Positive, Positive^+]$, and C_q is the target upper layer attribute where $C_q \epsilon C = \{C_1,\ C_2,\ \cdots,\ C_k\}$, in which C concludes all the defined categories. Therefore, the IF-THEN rule $E_{m+1}\ is\ L_{q(m+1)}\ and\ \cdots and\ E_n\ is\ L_{qn})$ can be extended as follows, where CW_q is defined to measure the weights for targeted attribute C_q by calculating all related association rules for selected sentiment categories in the set L_q:

$$\text{Rule } R_q{:}\ IF\ a_i\ is\ L_{q1}\ and\ \cdots and\ a_m\ is\ L_{qm}\ THEN\ Attribute\ C_q\ and\ CW_q$$

In order to distinguish this from the association rule mining problem, a newly defined single row of the training data LD, i.e. $X_p = \left\{x_{p1},\ x_{p2},\ \cdots,\ x_{pm}\right\}, p = 1, 2, \ldots, m,$ is used to mark the number of attributes, and define $\mu_{L_{q1}}\left(x_{p1}\right)$ as the sentiment possibility of data x_{p1} for one attribute for one sentiment category. Moreover, an operation is designed to summarize all the factors which are equal to the attributes from all the found interested association rules measured by sentiment set L_q as follows:

$$\mu_{L_q}\left(X_p\right) = \mu_{L_{q1}}\left(x_{p1}\right) \times \mu_{L_{q2}}\left(x_{p2}\right) \times \cdots \times \mu_{L_{qn}}\left(x_{pn}\right) \tag{9}$$

Hence, the confidence measure for the association rule $L_q \rightarrow C_q$ can be defined as follows:

$$CW_q = \mathrm{C}(L_q \rightarrow C_q) = \frac{\sum_{x_p \in C_q} \mu_{L_q}(X_p)}{\sum_{p=1}^{m} \mu_{L_q}(X_p)} \quad (10)$$

When the association rules have a single attribute supervised by one sentiment category, each association rule casts a vote for its consequent attribute. The strength of the vote is defined by the product of the compatibility grade and the rule weight. Hence, based on the confidence measured for each interested association rule the total strength of the vote for each attribute supervised by each sentiment category can be calculated as follows:

$$W_{Class\,h}(X_p) = \sum_{C_q = h} \mu_{L_q}(X_p) * CW_q \,; h = 1, 2, \cdots k \quad (11)$$

By using Formula (11), for each target upper layer attribute, all of the possibilities can be outputted with numeric values under each defined sentiment categories. The calculation results are presented as $W_{Class}(X_p) = \max(W_{Class\,1}(X_p),\ W_{Class\,2}(X_p),\ \cdots,\ W_{Class\,k}(X_p))$. For instance, for the selected top layer attribute "Apple.inc", the method will measure the weights of all the bottom layer attributes such as "CPU performance" and "appearance" by calculating all the relationships (interested association rules) from the bottom layer attributes to the top layer attribute. Overall, the proposed algorithm for handling the classification problem is to calculate a weight for the target attribute using one defined sentiment category by summarizing the weights of all the interested association rules where the consequent attribute is also the target attribute using one defined sentiment category. By taking into account all the related association rules, this method offers a solution to the problem identified, namely, how to handle the missing and embedded information from the social media data.

IV. Experiment

A. Experiment Environment

The Twitter dataset used for the experiment is a 1% sampling collected from Twitter.com by a web crawler system. It includes all the tweets from October 2011 to March 2012 which is huge data contains more than 200 million textual records, and stored it in MongoDB, which is a 'NoSQL' open source database in order to better processing the big data. In MongoDB, the data is stored in the JSON-style which offers both simplicity and power, and has extensive attributes for index support. The environment used for the experiment is the Operating System: Ubuntu12.04 X86_64 GNU/Linux with the CPU of Intel i7 3720QM and 16GB of RAM; and SUN Java jdk 1.7.0_25.

B. Predicting Stock Price Movements

We has designed an interesting problem to discover whether the stock price of some listed companies can be predicted by the sentiments expressed by the public. Guided by the Standard Industrial Classification (SIC), a standard system for classifying industries by a four-digit code, this study selected 30 highly topical listed companies in the NASDAQ and on the New York Stock Exchange, covering different industries with various products as the research targets. Almost 100 million records of Tweets which mention these companies or their products were retrieved. Using the collecting attribute "Geo", which refers to the location of the posted Tweets, only Tweets posted in the US were retained as the selected companies are all listed in NASDAQ Stock Market or on the New York Stock Exchange. More specifically to formulize this prediction problem, the hourly rise and fall of the stock price is normalized into five categories as up$^+$, up, flat, down and down$^-$ as shown in Table 2.

TABLE 2. HOURLY RISING AND FALLING CATEGORIES AS TRAGET ATTRIBUTE

up+	up	flat	down	down-
above +1%	+0.25% --- +1%	-0.25% --- +0.25%	-1% --- -0.25%	below -1%

	Daily Real Stock Price up or down			
	Hour_1	Hour_2	…	Hour_m
Company_1	up	down	…	up
Company_2	down-	flat	…	down
Company_3	flat	up	…	flat
Company_4	flat	flat	…	up
Company_5	up	up	…	up
...	…	…	…	…

Through the calculated tweets sentiment trends, we can first create a two-dimensional plot to show the brief correlation between public sentiment and the real stock price movements, as figure 3 shows. The ordinate means the numerical daily average sentiment degree, which ranges from -2 to 2, and the abscissa means the time period, which ranges from 1 to n. Figure 1 shows the fitted exampled companies' daily sentiment degree trend and its real stock price trend. The real stock price values in NASDAQ Stock Market and New York Stock Exchange were obtained using Yahoo! Finance and only considered the closing values.

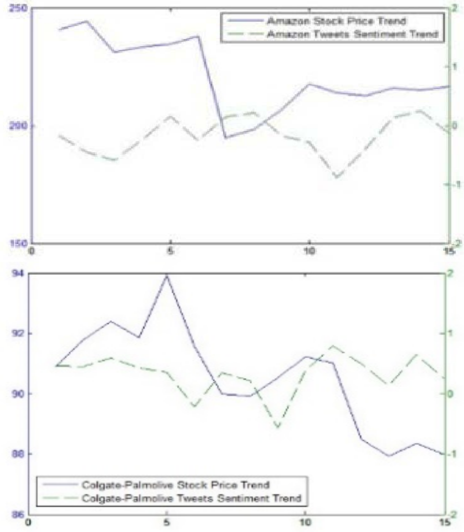

Fig. 3. Exampled Daily Fitted Trend for Amazon and Colgate-Palmolive

We further implement the proposed sentiment mining algorithm to use public sentiment trends to predict the stock price movements. To accompany the calculated sentiment possibilities for the bottom layer of attributes for the selected company, namely Apple Inc., the daily rising and falling categories is filled with values for each target attribute in column E_c as shown in Table 3.

TABLE 3. EXAMPLED ATTRIBUTES TABLE FOR ONE COMPANY

LD (Apple.Inc)		E_1	E_2	..	E_n	E_c
		iPhone_color	iPad_weight	..	iTunes_service	Stock price categories
S_1	Day_1	sentiment possibilities for $LD_{1,1}$	sentiment possibilities for $LD_{1,2}$	..	sentiment possibilities for $LD_{1,n}$	$V(E_{c1})$
S_2	Day_2	sentiment possibilities for $LD_{2,1}$	sentiment possibilities for $LD_{2,2}$	..	sentiment possibilities for $LD_{2,n}$	$V(E_{c2})$
…	…	…	…	..		
S_m	Day_m	sentiment possibilities for $LD_{m,1}$	sentiment possibilities for $LD_{m,2}$	..	sentiment possibilities for $LD_{m,n}$	$V(E_{c3})$

It is reasonable to expect that public opinion on companies will only affect their stock prices with some delay. Accordingly, the influence of the time period on the share price is also considered in this study. As shown in Table 4, T+X in the first row uses the current sentiments expressed in Tweets to predict the rise and fall of the stock price X days later. The results of the experiment how that in general, for the 30 representative listed companies, the average predictive accuracy from T+3 was the highest, yielding up to 66.48%. Moreover, among all the different industries, if using the T+3 time slot for the prediction, the data shows that the IT industry has the highest predictive accuracy, achieving 76.12%. In contrast, predictions for manufacturing companies are the least accurate, viz., 52.94%. Table 5.3 also suggests another interesting factor, i.e., a delay match between the trading date and the settlement date. In a stock market, the trading date means the day when a customer's trade is executed. Once the trade is placed, the customer's order will go to the stock exchange and a trade confirmation will be posted to the customer's account immediately. However, some time is still needed to clear the transaction and transmit the funds from the buyer to the seller. So the settlement date is the day when this operation is completed. Under the regulations of the American stock market, the settlement date is 3 days after the trade date, and this is almost consistent with the hypothesis of the best predictive accuracy in this study. Figure 4 shows a graph of the normalized stock up or down as predicted by our proposed algorithm vs the actual stock price trend when trained on T+3 time slot for the listed company Amazon as example, and we find the correlation is very close.

TABLE 4. RESULTS ON PREDICTIVE ACCURACY

T+0	T+1	T+2	T+3	T+4
57.42%	59.38%	61.29%	66.48%	61.35%

IT	Finance	Manufactory	Medicine	Media	Energy
76.12%	70.75%	52.94%	61.89%	73.78%	63.37%

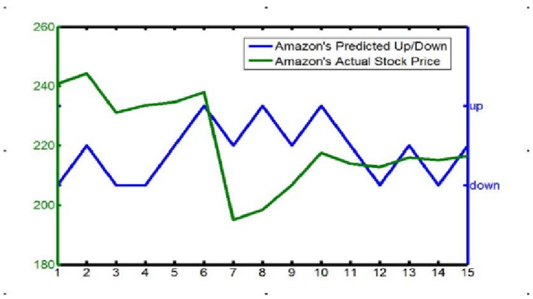

Fig. 4. Correlation between the Actual Stock Price and the Predicted up or down for Amazon

We also implement other three classic data mining algorithms for our experimental dataset to compare the predict accuracy with our proposed algorithm, which are Naïve Bayers classifier, support vector machine (SVM) and C4.5. Compared with other classic data mining algorithms, as table 5 shows, proposed algorithm achieves the best average result. Only the proposed algorithm, SVM has better average predictive accuracy. The nature of the experimental data strongly influences the performance of an algorithm. For instance, Apriori has the lowest predictive accuracy since it has to deal with noisy data and Tweets always contain a large amount of ambiguous data. Lacking the capacity to handle multi-layered attributes, C4.5 also performs badly as there is a considerable proportion of missing values for some attributes due to the nature of social media data. This may affect the accuracy by mistakenly ignoring some bottom layer information even if the upper layer of attributes is empty.

TABLE 5. COMPARISONS WITH OTHER ALGORITHMS WITH SELECTED INDUSTRIES AND OVERALL AVERAGE ACCURACY

	Proposed Algorithm	SVM	C4.5	Naïve Bayers
IT	76.12%	70.75%	67.63%	65.02%
Manufactory	52.94%	55.32%	51.48%	53.13%
Media	73.78%	71.81%	63.84%	62.22%
Average	66.48%	63.34%	60.15%	59.79%

V. CONCLUSION

Through the evaluation about recent social media analysis works especially on public opinion mining up-to-date, this study has identified and emphasized two important research problems. Through the proposed whole methodology for social media data mining, this study has fulfilled the identified research problems with satisfactorily experimental results. More specifically, the proposed algorithm has defined the whole relationships embedded in social media as a graph with several layers, which the top layered attributes and intermediate layered attributes have direct relations (belongs), and the bottom layered attributes and intermediate layered have indirect relations (describes). Combined with both direct and indirect division operations accompany with applied the chi-square test and adjusted residual, the proposed process has increased the dimensionality of whole attributes that could significantly measures the missing and embedded information among the social media data, thus to achieve better predict accuracy compared with other existed direct social media mining method through the experiment results.

Furthermore, this research has two major practical implications. On one hand, the proposed algorithms have a better prediction performance in some certain industries such as IT and media. This knowledge should aid these companies to effectively manage or promote products and brands via sentiment management in social media. On the other hand, our study indicates the proposed algorithms have a better performance in using current tweets' sentiment to predict the stock price of three days later. This knowledge informs that a 3-day interval is the best period to evaluate the efficacy of event management in social media.

At the same time, this study has opened up several opportunities for future research. Firstly, our research only considers the daily or weekly closing values of the stock price. However, for the purpose of prediction and obtaining actual investment incomes, the opening, highest and lowest values of a time series have the equal importance as the closing values. The future research on include these values into the analyses. Secondly, according to our experimental findings, using current public sentiment to predict real stock market price with a period of 3 days (i.e., T+3 days) has the best predict accuracy. While in the procedure of time divisions, we only selected the trading days for our analysis but ignoring those non trading days like holidays and weekends. We believe in those non trading days, the public sentiments keep accumulating and therefore can cause a timeline gap in our proposed algorithm. These problems may be potentially addressed by increasing the complexity of input stock market data as the prediction target, adding more attributes in the analyses and enriching our proposed algorithm. Moreover, the dataset which we used is the Twitter data. Most Twitter messages are very short and some of them are actually meaningless. As we calculated daily sentiment degree of selected listed companies based on each record of tweet, the very short and meaningless information has reduced the accuracy of the sentiment classification algorithm and the accuracy of stock market price prediction. To address this problem, we suggest adding other social media data, like those from Facebook, which may include long textual data in order to further the prediction accuracy.

References

[1] E. Schonfeld. Mining the thought stream. TechCrunch Weblog Article. 2009.

[2] http://blog.twitter.com/2011/03/numbers.html.

[3] J. Leskovec, L. Adamic and B. Huberman. The dynamics of viral marketing. In Proceedings of the 7th ACM Conference on Electronic Commerce. 2006.

[4] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology. 2009.

[5] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology. 2009.

[6] D. Gruhl, R. Guha, R. Kumar, J. Novak and A. Tomkins. The predictive power of online chatter. SIGKDD Conference on Knowledge Discovery and Data Mining. 2005.

[7] M. Joshi, D. Das, K. Gimpel and N. A. Smith. Movie Reviews and Revenues: An Experiment in Text Regression NAACL-HLT. 2010.

[8] S. Asur and A. Huberman. Predicting the Future with Social Media. CoRR, abs/1003.5699. 2010.

[9] J. Bollen, H. Mao and Z. Xiao-Jun. Twitter mood predicts the stock market. Journal of Computational Science, 2, 1-8. 2010.

[10] T. Leung, H. Daouk and A. Chen. Forecasting stock indices: a comparison of classification and level estimation models. International Journal of Forecasting, 16, 173–190. 2000.

[11] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology. 2009.

[12] D. Gruhl, R. Guha, R. Kumar, J. Novak and A. Tomkins. The predictive power of online chatter. SIGKDD Conference on Knowledge Discovery and Data Mining. 2005.

[13] M. Joshi, D. Das, K. Gimpel and N. A. Smith. Movie Reviews and Revenues: An Experiment in Text Regression. Proceedings of NAACL-HLT. 2010.

[14] R. Sharda and D. Delen. Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications, 30, 243–254. 2006.

[15] S. Asur and A. Huberman. Predicting the Future with Social Media. CoRR, abs/1003.5699. 2010.

[16] D. M. Pennock, S. Lawrence, C. L. Giles, and F. A. Nielsen. The real power of artificial markets. Science, 291, 987–988. 2001.

[17] W. Zhang and S. Skiena. Improving movie gross prediction through news analysis. Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. In Web Intelligence, 1, 301-304. 2009.

[18] R. P. Schumaker and H. Chen. Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System. ACM Trans. Inf. Syst. 27, 1–19. 2009.

[19] A. Mittal and A. Goel. Stock Prediction Using Twitter Sentiment Analysis. 2011.

[20] Chan, C. C., Wong, K. C. & Chiu, K. Y. Learning Sequential Patterns for Probabilistic Inductive Prediction. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, 24 (10), 1532-1547. 1994.