

Survey : Sentiment Analysis of Twitter Data for Stock Market Prediction

Rupawari Jadhav¹, M. S. Wakode²

Department of Computer Engineering, Pune Institute of Technology, Pune ^{1,2}

Abstract: Stock market prediction is trying to determine the future value of a company stock. The company should make a profit when the prediction of stock's future price could be successfully done. News article and social media can affect the stock market. The Twitter is one of the most useful social media that can help to predict public mood. The predicted mood and previous day value are useful for predicting stock market price. This paper discusses the prediction of the future stock market with help of sentiment score. The different techniques can be used to classify result of sentiment score. In this paper, the different technique for stock market prediction is introduced. We also discuss technology that speedup the computation which will improve the performance.

Keywords: Clustering, classification techniques, Sentiment analysis, Social data analysis, Stock market prediction.

I. INTRODUCTION

Stock Prices are considered to be very dynamic and impressible to quick changes because of the underlying nature of the financial domain. Researcher in financial domain conclude that the news article, blogs, and stock market prediction is important topics in many business income. Stock market forecasters focus on developing a successful approach for forecast the values or stock prices. The stock market prices are more fluctuating that's fall the stock price or raising the stock price. The aim is to earn high profit using well defined different trading strategies. The overall mood of social data according to the company that can be important variables which affect the stock price of the company. The different on-line social network sites that helps to make availability of large amount of data. Therefore, comparing information from social media data with historical price can improve the prediction and accuracy of a model.

Twitter is currently the most popular social microblogging platform [1] allowing its users to send and read short messages of up to 140 characters in length, known as tweets. It provide different services to create and share massive amount of data. From Twitter analysis of company's product can improve relationship and trust between customer and producer according to need of customer.

Sentiment analysis is a study that addresses opinion-oriented natural language processing. Such opinion oriented studies include, among others, emotion and mood recognition, ranking, relevance computations, related in text [7]. G-POMS, Google N-gram, Lingmotif, LIWC, POMS, Opinion Finder, SentiStrength are the different analytical tools provide to calculate sentiment analysis of given data.

The problem definition is to develop a model for sentiment analysis within big data distributed environment for stock prediction. To apply clustering and SVM classifier on sentiment score to improve accuracy and

implement the model in distributed environment to speeds up the performance. The dataset must be filtered by adding metadata such as exact location of a person, the number of re-tweets, the number of followers to selected data-set. To parallelize the computation using Map-reduce distributed environment.

II. LITERATURE SURVEY

Sentiment analysis is the most important concept of the research area in various different fields. There are many researchers that studies and aim to identify a method to predict sentiment analysis on different area fields. Social media is popularized source of data which collect useful data such as blogs, micro-blogs, Facebook, Twitter etc.

The paper[1] covers the evaluation of a system that can be used to predict future stock price based on analysis of social media data. Twitter messages are retrieved in real time using Twitter Streaming API. The large volume of data to be classified using Naive Bayes method for fast training process with a large volume of training data. The stock market prediction should be calculated by using linear regression technique.

The paper[2] presents the two different textual representations, Word2vec and N-gram, for analyzing the public sentiments in tweets. The author applied sentiment analysis and supervised machine learning principles (such as logistic regression, random forest, SMO) to tweets extracted from twitter and analyzing the correlation between stock market movement of company and sentiments in tweets. A data can be extracted from twitter API of Microsoft using keyword \$MSFT, #Microsoft, etc.

In this paper[3], the authors created a system that predicts stock market movements on a given day, based on time series data and market sentiment analysis. They collect prices for S&P 500 from January 2008 to April 2010 from Yahoo! Finance into Excel spreadsheet. For

sentiment analysis, they obtained Twitter Census stock Tweets data-set from Info-chimps, a privately held company that offers a “data marketplace”. Naive Bayes Classifier used to analyze sentiment in the tweet data set. The SVM, Logistic and Neural network techniques would be used for predicting market movement.

The paper[4] focuses on the prediction of daily stock movements of three Indian companies listed on National Stock Exchange (NSE). The Support Vector Machine (SVM) was used for prediction of the stock market. The tweets collected were of 5 month period having 200000 tweets. The tweets were collected directly from twitter using Twitter API and filtered using keywords for example #airtel. The relevant stocks were downloaded directly from yahoo finance.

In this paper[5], sentiment analysis of a product is performed by extracting tweets about products and classifying the tweets that can be as positive and negative sentiment. This paper proposes a hybrid approach which combines unsupervised learning to cluster the tweets and then performing supervised learning methods for classification. In this paper, 1200 tweets were collected for the company ‘Apple’ for analysis. The proposed model would be compared with SVM, CART, Random forest, Logistic regression. The predicted and actual value can be compared using confusion matrix.

In this paper[6], the author examines the effectiveness of various machine learning techniques on providing a positive or negative sentiment on a tweet. The author applies different machine learning techniques: Naive Bayes, Maximum entropy, support vector machine etc. and compare them. They looked for a correlation between twitter sentiments with stock prices and determined which words in tweets correlate to change in stock price by doing a post analysis of price change and tweets.

From the literature survey, we can conclude that for sentiment analysis of bigger data-set made to be accurate and efficient we need to make use of distributed approach. In this paper, we introduce a distributed model with supervised and unsupervised technique to improve accuracy and performance.

III. DATA SET

A. Social Media :

The following features capture useful aspect of Twitter and authors for opinion retrieval[17].

1. **URL** : Most Tweets containing a link usually give the objective introduction to the links. Additionally, spam in Twitter often contain links. Hence, we use a feature indicating whether a Tweet contains a link in our ranking model.
2. **Mention** : In a Tweet, people usually use “@” preceding a user name to reply to other users. The text of this Tweet is more likely to be ‘personal content’.
3. **Hashtag** : A hashtag refers to a word in the text

of the Tweet that begins with the “#” character. It is used to indicate the topic of the Tweet.

4. **Recency** : Twitter generates streams of text in real time and it is often hypothesized that more recent results are better for Twitter retrieval.

Twitter is a social network. The more author information can also be used for the analysis of spammer detection.

1. **Statuses** : The number of Tweets (statuses) the author has ever written that which related to the activeness of an author. The most active authors are likely to be spammers who post very large number of Tweets. Therefore, we use the number of statuses as a feature for Tweets ranking.

2. **Place** : The location associated with tweet that help to identify which place the tweet should be posted. In this contains country, country code, id, name of the place.

3. **Followers and Friends** : In Twitter a user can choose to follow any number of other users that he finds interesting for one reason or another. If userA follows userB, all the Tweets posted by userB will be updated in the userA’s private stream. We call userA a follower of userB and userB a friend of userA. The number of followers indicates the popularity of the user. The number of friends also reflects the type of the user.

4. **Retweets** : Twitter provide services to user to retweet tweets generated by other user. All retweets start with symbol indicated by @RT. The retweet of the most recent tweets of a user is also one of the feature in spam detection system.

5. **Listed** : A user can group their friends into different lists according to some criteria. If a user is listed many times, it means that his Tweets are interesting to a large user population. We use a feature that measures how many times the author of a Tweet has been listed for Tweet ranking.

B. Classification Techniques:

Supervised learning is an important technique for solving classification problems in sentiment analysis. Training and testing the data make it more easier for prediction future data. The different classifiers can be used to classifies the sentiment score of each tweets that predict the emotion of the tweets. The different techniques are compared as shown in Table I [16] with their result and accuracy.

1. Naive Bayes :

Naive Bayes is a conditional probability model that given a problem instance to be classified. Naive Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variable.

$$P(c|d) = \frac{(P(c) \sum_m^{i=1} P(f_i|c)P(f_i|c)^{n_i(d)})}{P(d)}$$

In this formula, represents a feature and $n_i(d)$ represents the count of feature f_i found in tweet d . There are a total m features. Parameters $P(c)$ and $P(f_i|c)$ are obtained through maximum estimates [1][15].



2. SVM :

Vector machines (SVM) are primarily Classifiers that can classify by constructing hyperplanes that separate cases that belong to different categories. A Support Vector Machine (SVM) is a supervised classification algorithm that recently has been applied successfully to text classification tasks.

$$c(x) = \begin{cases} 1 & w \cdot \phi(x) + b \geq k \\ -1 & w \cdot \phi(x) + b \leq -k \end{cases}$$

where, $w = \{w_1, \dots, w_n\}$ is a weight vector.

$x = \{x_1, \dots, x_n\}$ is a input vector.

$\Phi(x)$ is kernel function.

Among these two classifier it is observed that SVM classifier outperforms every other classifier in predicting the sentiment of the tweets[11][14].

3. Maximum Entropy :

The Maximum entropy Classifier converts labeled feature sets to vectors using encoding. This encoded vector is then used to calculate weights for each feature that can then be combined to determine the most likely label for a feature set[11]. The fundamental concept of maximum entropy is that if much information regarding the data is not known, the distribution should be extremely uniform[17].

4. Random Forest:

Random Forest classifier is a tree-based classifier. It consists of numerous classification trees that can be used to predict the class label for a given data point based on the categorical dependent variable[17]. The error rate of this classifier depends on the correlation among any two trees in the forest that adds the strength of definite or individual tree in the forest. In order to minimize the error rate, the trees should be strong and the degree of correlation should be as less as possible.

TABLE I Comparative Results for Sentiment Classification Techniques

Model/ Algorithm	Dataset	Accuracy (%)
SVM	Amazon product review data and ChnSentiCorp dataset	89.8
NB	Amazon product review data and ChnSentiCorp dataset	89.4
NBSVM	SEMEVAL 2013 twitter dataset	79.4
MNB	SEMEVAL 2013 twitter dataset	71.14
SentiStrength+TwitterSentiment	SNAP twitter dataset	69.7
SentiStrength	MySpace dataset	62.3
TwitterSentiment	SNAP twitter dataset	57.2

The Table II defines the different parameter of supervised machines learning algorithm that can be used for classifying the sentiment of twitter data[17].

TABLE II PARAMETRIC COMPARISON OF THE SUPERVISED MACHINE LEARNING ALGORITHMS

ALGORITHM	NB	SVM	MAXIMUM ENTROPY	RANDOM FOREST
COMPLEXITY	VERY LESS	HIGH	MODERATE	MODERATE
THEORETICAL ACCURACY	LOW	HIGH	MODERATE	HIGH
THEORETICAL TRAINING SPEED	HIGH	HIGH	MODERATE	LOW
PERFORMANCE WITH SMALL NO. OF OBSERVATIONS	HIGH	LOW	LOW	LOW
CLASSIFIER	PROBABILISTIC	LINEAR	PROBABILISTIC	TREE BASED

C. Prediction Techniques :

Depending upon the share that should be rise or fall stock market analysis can be classified as[18] :

- fundamentalistic that consider only fundamental factor of the stock market.
- chartists, they use graphical design.

The fundamental analysis uses mathematical, statistical, and financial algorithms, applied to the periodic financial statements of the company, in order to evaluate, the shares' price as correctly as possible. The technical analysts study the short-term changes in the shares' price,

with the study of the history of the quotations and assumption of the past behavior will extend into the future. The technical analysis provide information about the possible future evolution of the stock market.

A. Method of Time series Analysis :

The time series model make consideration of the past behavior of a given variable and uses this information that helps to predict its future behavior. A time series model will be referenced when: (1) we have less information on the factors that affect the behavior of the variable, (2) we have a huge amount of data, or (3) the main aim is short-term prediction of the stock[18].



1. Random Walk :

The random walk theory suggests that stock price changes have the same distribution and are independent of each other. So that the past movement or trend of a stock price cannot be used to predict its future movement. The stock price are fluctuating and status of financial fields of market can be predicted as random walk. It is simple to use and it can easily handle flows around complicated boundaries. It does not exactly because it is impossible to consistently outperform the market, in the short-term. Next, the computed outcome are noisy due to the statistical errors.

2. Moving Average :

It is widely used indicator in technical analysis that helps smooth out price action and filter out the noise from random price fluctuations. It is based on the past price of stock market. There are two commonly used MAs are the simple moving average (SMA), which is the simple average of a security over a defined number of time periods, and the exponential moving average (EMA), which gives bigger weight to more recent prices. The quick turning markets stock price will always be well ahead of the moving averages[18].

3. Regression method :

Linear regression is most common technique for predicting the future value of variable based on linear relationship it has with other[18]. It assume one straight line that approximates the given data, and forecast the future value based on direction of the regression line[19].

The equation(1) that describes relationship of dependent and independent variables with straight line:

$$y = a + bx$$

where, y is the dependent variable,
x is the independent variable,
a and b are line's coefficient.

The validation of the linear regression model summarize using R^2 . This can be calculated as the square of the correlation between the observed values and the predicted values.

4. ARIMA model :

ARIMA models are applied in that cases where data show evidence of non-stationarity also where an integrated part can be applied to eliminate the non-stationarity. Its main application is in the area of short term forecasting require at least 38 historical point. It works best when your data exhibits a stable or consistent pattern over time with a minimum amount of outliers. ARIMA is usually best to exponential smoothing techniques when the data is moderately long and the correlation between past observations is stable[18].

B. Evaluation of prediction method :

The predicted error factor should be considered to calculate accuracy of different prediction algorithm. The Error means that the difference between the predicted

and actual dataset.

For a total of N given historical data in the dataset, p_x and the corresponding forecasted or predicted value \hat{p}_x , such that $x = 1$ to N, the performance evaluation factors are given as [20],

- Mean Absolute Error :

$$MAE = \frac{1}{N} \sum_{x=1}^N |p_x - \hat{p}_x|$$

- Normalised Mean Absolute Error :

$$NMAE = \frac{MAE}{p_{max} - p_{min}}$$

- Mean Absolute Percentage Error :

$$MAPE = \frac{1}{N} \sum_{x=1}^N \left| \frac{p_x - \hat{p}_x}{p_x} \right| * 100$$

- Mean Square Root Error :

$$MSRE = \frac{1}{N} \sum_{x=1}^N \sqrt{(p_x - \hat{p}_x)^2}$$

- Mean Square Error :

$$MSE = \frac{1}{N} \sum_{x=1}^N (p_x - \hat{p}_x)^2$$

- Root Mean Square Error :

$$RMSE = \sqrt{\frac{1}{N} \sum_{x=1}^N (p_x - \hat{p}_x)^2}$$

IV. CONCLUSION

This paper study on the survey of different methods that gave helps the prediction of stock market using financial news from Twitter data. It also discussed different classification algorithm for sentiment analysis and the comparison between them. The twitter data contains spam tweets that also identified and removed on the basis of factor that helps to detect spam. The different error technique also defined that helps to calculate the error of different prediction model.

REFERENCES

- [1] Michal Skuza, Andrzej Romanowski, "Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction", Computer Science and Information Systems pp. 1349–1354, 2015 F230 ACSIS, Vol.5.
- [2] Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, Babita Majhi, "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements", International conference on Signal Processing, Communication, Power and Embedded System (SCOPEs), 2016.
- [3] Tina Ding, Vanessa Fang, Daniel Zuo, "Stock Market Prediction based on Time Series Data and Market Sentiment", 2012.
- [4] Phillip Tichaona Sumbureru, "Analysis of Tweets for Prediction of Indian Stock Markets", International Journal of Science and Research (IJSR), Volume 4 Issue 8, August 2015.



- [5] Rishabh Soni, K. James Mathai, "Improved Twitter Sentiment Prediction through 'Cluster-then-Predict Model'", International Journal of Computer Science and Network, Volume 4, Issue 4, August 2015.
- [6] Linhao Zhang, "Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation", April 16, 2013.
- [7] Manoj Kumar Danthala, "Tweet Analysis: Twitter Data processing Using Apache Hadoop", International Journal Of Core Engineering & Management (IJCEM) , Volume 1, Issue 11, February 2015.
- [8] www.tutorialspoint.com/hadoop/pdf/hadoop-mapreduce.pdf.
- [9] Phayung Meesad, Jiajia Li, "Stock Trend Prediction Relying on Text Mining and Sentiment Analysis with Tweets", 2014.
- [10] Ms. Gaurangi Patil, Ms. Varsha Galande, Mr. Vedant Kekan, Ms. Kalpana Dange, "Sentiment Analysis Using Support Vector Machine", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, January 2014.
- [11] Alec Go, RichaBhayani, Lie Huang. , "Twitter Sentiment Classification Using Distant Supervision", CS224N Project report, Stanford, Google scholar, Page1-12, 2009.
- [12] Manoj Kumar Danthala, "Tweet Analysis: Twitter Data processing Using Apache Hadoop", International Journal Of Core Engineering & Management (IJCEM), Volume 1, Issue 11, February 2015.
- [13] Abinash Tripathya, Ankit Agrawalb, Santanu Kumar Rathc, "Classification of Sentimental Reviews Using Machine Learning Techniques", International Conference on Recent Trends in Computing 2015, Procedia Computer Science 57 (2015) 821 – 829.
- [14] Jose, A.K., Bhatia, N., Krishna, S. : "Twitter sentiment analysis", Major Project Report, NIT Calicut (2010).
- [15] Varsha Sahayak, Vijaya Shete, Apashabi Pathan, "Sentiment Analysis on Twitter Data", International Journal of Innovative Research in Advanced Engineering (IJIRAE), Issue 1, Volume 2 (January 2015).
- [16] Bholane Savita Dattu, Prof.Deipali V. Gore , "A Survey on Sentiment Analysis on Twitter Data Using Different Techniques", International Journal of Computer Science and Information Technologies, Vol. 6 , 2015.
- [17] Mitali Desai , Mayuri A. Mehta, "Techniques for Sentiment Analysis of Twitter Data: A Comprehensive Survey", International Conference on Computing, Communication and Automation (ICCCA2016).
- [18] Virginica Rusu and Cristian Rusu, "Forecasting methods and stock market analysis", 2005.
- [19] Dr. P. K. Sahoo, Mr. Krishna charlapally, "Stock Price Prediction Using Regression Analysis", International Journal of Scientific & Engineering Research, Volume 6, Issue 3, March-2015.
- [20] G.Mahalakshmi, Dr.S.Sridevi, Dr.S.Rajaram, "Survey on Forecasting of Time Series Data ", IEEE, 2016.

BIOGRAPHIES

Rupawari A. Jadhav is a student in the Computer Engineering Department, College of Pune Institute of Technology, Pune University. She received Master of Computer Engineering (ME) degree in 2017 from P. I. C. T., Pune, India.

M. S. Wakode is a professor in the Computer Engineering Department, College of Pune Institute of Technology, Pune.