International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:9, No:2, 2015

Semantic Enhanced Social Media Sentiments for Stock Market Prediction

K. Nirmala Devi, V. Murali Bhaskaran

***Abstract*—**Traditional document representation for classification follows Bag of Words (BoW) approach to represent the term weights. The conventional method uses the Vector Space Model (VSM) to exploit the statistical information of terms in the documents and they fail to address the semantic information as well as order of the terms present in the documents. Although, the phrase based approach follows the order of the terms present in the documents rather than semantics behind the word. Therefore, a semantic concept based approach is used in this paper for enhancing the semantics by incorporating the ontology information. In this paper a novel method is proposed to forecast the intraday stock market price directional movement based on the sentiments from Twitter and money control news articles. The stock market forecasting is a very difficult and highly complicated task because it is affected by many factors such as economic conditions, political events and investor's sentiment etc. The stock market series are generally dynamic, nonparametric, noisy and chaotic by nature. The sentiment analysis along with wisdom of crowds can automatically compute the collective intelligence of future performance in many areas like stock market, box office sales and election outcomes. The proposed method utilizes collective sentiments for stock market to predict the stock price directional movements. The collective sentiments in the above social media have powerful prediction on the stock price directional movements as up/down by using Granger Causality test.

***Keywords*—**Bag of Words, Collective Sentiments, Ontology, Semantic relations, Sentiments, Social media, Stock Prediction, Twitter, Vector Space Model and wisdom of crowds.

I. Introduction

PREDICTION of the stock price directional movement has been one of the hot research topics for decades. Recently, several attempts have been created to enhance the accuracy of the predictions by knowledge from social media like Face book, Google and Twitter. Data from social media are important indicators for sentiments that doubtless carry helpful information in addition to financial content.

The stock market forecasting is a very difficult and highly complicated task because it is affected by many factors such as economic conditions, political events [4] and investor's sentiment etc. The stock market series are generally dynamic, nonparametric, noisy and chaotic by nature. Current prediction strategies, however, not provided the results that are easily interpretable. To obtain interpretable results, regression methods that induce scantiness are required and information that's not helpful for the prediction is automatically discarded from the model. The social media are platforms of open, honest and real sharing of news to clarify investment behavior in the stock market.

Nowadays, the tremendous growth of information is available in social media sites such a forums [17], twitter [3], [10], [11], [13], [14], [20], Facebook, blogs and news reports etc., are having large volume of public opinion information. It is essential to extract sentiment information from these social networks for making predictions in time and understand the trends of the opinion correctly. Exploiting social media sentiment textual information in addition to numeric historical time series stock data increases the prediction accuracy with quality of the data.

Sentiment Analysis is a process of obtaining the attitude / opinion of authors about particular objects. Recently, due to the tremendous growth of sentiment information from various social medias, sentiment analysis has become one of the promising research areas in computational linguistic. The growing importance of Sentiment Analysis applied to finance brings forth several analysis and sensible problems to minds. In finance, there have been numerous studies using textual analysis to compute the sentiment of diverse news things, articles, monetary reports, and tweets regarding public corporations. Then, the examined sentiments are often accustomed replicate the correlations with alternative monetary measures, like stock returns and volatilities.

The 'wisdom of crowds' equipped with text mining and sentiment analysis will automatically generate collective intelligence of future performance on a numerous areas such as sports outcome, hotspot forums prediction, stock market price prediction, election results and box office sales prediction. Over the past few years, important progress has been achieved in exploitation Twitter as a further supply of knowledge. The sentiment lexicon is the important resource in many sentiment oriented applications. Tim Loghran and Bill McDonald [2] proposed a common Psycho sociological dictionary with extended finance specific lexicons.

The primary aim of this research paper is to investigate the predictive power of social media sentiments in stock price directional movement. Experimental results show that there is some causal relationship between public sentiment and stock market indices to provide useful investment decisions in the right direction. The accuracy of stock price directional movements can be significantly improved by the incorporation of semantic enhanced sentiments from the social media.

The rest of the paper is structured as follows: Section II briefly discusses the review of related works. The details of the proposed system are presented in Section III. The experimental setup and empirical results drawn from the

K.Nirmala Devi is with the Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India (e-mail: sunsys19@yahoo.com).

Dr. V. Murali Bhaskaran is with the Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India (e-mail: murali66@gmail.com).

proposed system is compared with the others is discussed in Section IV and Section V concludes the paper.

II. Review of Related Works

The Efficient Market Hypothesis (EMH) and Random Walk were mostly used in the early stock market predictions. However, the growing research critically examined EMH from the perspective of behavioral economics. Many studies shown that stock market prediction do not follow a Random Walk and will indeed to some degree to be predicted.

The movie sales prediction was also done with blog sentiments. Liu et al. [21] proposed a model based on Probabilistic Latent Semantic Analysis (PLSA) to extract the sentiment indicators from blogs. The relationship between financial news and stock price movements also investigated. Although news certainly influences the stock market prediction and play a significant role in financial decisions. The sentiments expressed in twitter can predict the box office receipts. It is therefore reasonable assumption that the public sentiments in social media can predict the stock price directional movements as up/down.

Usually, the stock market prediction is to be done with either technical or fundamental indicators. The technical indicators are quantitative measure and are obtained from the historical data such as simple moving average, exponential moving average etc. The fundamental analysis is to be performed with the non historical quantitative information such as macroeconomic indicators and the majority of the data is of unstructured nature. Therefore, it is essential to extract the information from unstructured sources and perform the analysis to make use of them in the prediction work.

Yet, major works done with sentiment analysis mainly focuses on the product review, movie review and blogs [21]. On the other hand, the lexicon developed for one domain misclassifies information in another domain. Since, they are domain dependent.

Various kinds of forecasting models based on soft computing [5]-[9], [12], [16], [19] have been proposed to improve prediction accuracy in the stock market forecasting. Shom Prasad Das and Sudarsan Padhy [1] have developed a model based on Back Propagation Technique (BP) and Support Vector Machine Technique (SVM) to predict futures prices traded in Indian stock market. But Back Propagation lack convergence in learning. The performances of these techniques are compared and it is observed that Support Vector Machine (SVM) [7], [17], [18] provides better performance results as compared to Back Propagation (BP) technique.

Furthermore, the investors not aware much of the stock market behavior and they do not know which stock investment yields more profit. To know the stock market progress they want to analyze all relevant information from the news sources and magazines. The natural language processing and machine learning are playing major role in the prediction of the stock price.

Likewise, face book and Twitter [3], [10], [11] are most popular social media and has high influence in the stock market prediction. The content of the Twitter was used to predict the stock market movement in the Dow Jones [3], [5] as well as Indian stock index. Based on the list of six different states of mood such as calm, alert, sure, vital, kind and happy were used in the analysis. The authors found that happy and calm had a high correlation with the stock market prediction.

Using Naïve Bayeisan (NB) [15] approach the movement of the stock price was done based on news article's sentiment from micro blogging data to perform stock market forecasting has already presented promising results.

Mao et al. [14] used a random sample of public tweets with a sentiment to decide the tweet as "bullish" and "bearish" the stock market. They showed that their sentiment indicators and frequency of financial terms are significantly predicting the stock market returns. The bag of words approach was combined with J48 [16], [17] to predict the stock market movement based contents.

The proposed system made an attempt to investigate the predicting power of the collective sentiments from Twitter and money control news article in stock price prediction. The objectives of this paper is

- Determine the relationship between the trade volume of the stock and the user's discussion of stocks in the Twitter and money control news articles.
- Analyze whether users' collective sentiments of tweets have predictive power on the stock price directional movement as up/down.

III. The Proposed System

The tremendous growth of web arouses much attention on public opinion. It provides a lot of opportunities to analyze the public opinion. Stock markets are a major component of the world economy since they provide a large platform for companies to raise money efficiently and effectively. The schematic process of proposed system is shown in Fig. 1.

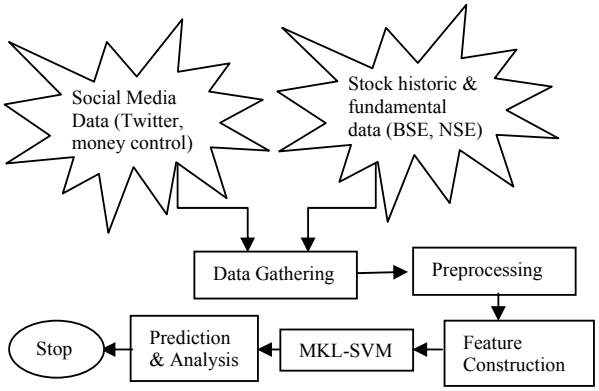

Fig. 1 The Proposed System Process

A. Data Gathering

The discussion forums and blogs are not rapidly progressing to spread the data but stock market requires the communication medium as rapidly spreadable nature. Rather than forums and blogs, the news channels and twitter are emerging their progress in rapid manner. So the proposed

system gathers the data from Twitter and money control news sources.

The data used in the proposed system consists of daily closing prices of Infosys, HCL, TCS, Tech Mahindra and Wipro from BSE, NSE and Yahoo finance [23].For fundamental analysis the closing prices of Gold, Exchange Rate (US$ to INR) and Crude Oil Price (COP) also extracted from BSE, NSE and Yahoo finance [23]. The sentiment data was collected from Twitter and money control [23] (India's Leading financial source). The corpus was constructed with all the above sources from January 2014 to July 2014.

B. Preprocessing

The raw data collected from the social media are preprocessed by eliminating the non opinion information, removal of stop words and stemming.

C. Feature Construction

The proposed system combines the indicators from technical, fundamental as well as from sentiment features. The technical and fundamental features are derived from historic data and sentiment features are extracted from social media. The system uses the five derived technical indicators from stock historic data (High, Low, Open, Close and Volume) and one output represent the next day's closing price as up / down / no change. The various technical indicators are calculated from the historical stock data as follows.

1. Relative Strength Index (RSI)

A technical indicator to determine over bought and over sold conditions of an asset based on recent gains and recent losses. The formula for RSI is calculated using (1) and (2).

$$RSI = 100 - \left[\frac{100}{1+RS}\right] \quad (1)$$

where

$$\text{RS} = \text{Average}\left[\frac{\text{x day's Up closing Price}}{\text{x day's Down closing price}}\right] \quad (2)$$

2. Money Flow Index (MFI)

The strength of money is measured by this indicator and the formula used for MFI is calculated using (3)-(5).

$$MFI = 100 - \left(\frac{100}{(1+Money\ Ratio)}\right) \quad (3)$$

where

$$\text{Money Ratio} = \frac{\text{Positive Money Flow}}{\text{Negative Money Flow}} \quad (4)$$

$$\text{Money Flow} = \text{Typical Price} * \text{Volume} \quad (5)$$

3. Exponential Moving Average (EMA)

The indicator of exponential moving average over a particular period is calculated using (6).

$$EMA = [\,\alpha * Today's\ Closing\ Price\,] + [\,1 - \alpha * Yester\ day's\ Closing\ Price] \quad (6)$$

4. Stochastic Oscillator (SO)

This is the measure of the difference between the current closing price of a security and its lowest low price, relative to its highest high price for a given period of time. The calculation is done with (7):

$$\%K = \left[\frac{(Closing\ Price - Lowest\ Price)}{(Highest\ Price - Lowest\ Price)}\right] * 100 \quad (7)$$

5. Moving Average Convergence Divergence (MACD)

This indicator finds the difference between a short and a long term moving average. It is calculated with its signal using (8) and (9).

$$MACD = [\,0.075 * E\,] - [\,0.15 * E\,] \quad (8)$$

where, E is EMA (Closing Price)

$$Signal\ Line = 0.2 * EMA\ of\ MACD \quad (9)$$

6. Stock Return:

The stock return for all the stocks is calculated using (10):

$$Stock\ Returns\ (SRET) = \ln\left[\frac{CP_t - CP_{t-1}}{CP_{t-1}}\right] \quad (10)$$

where CP_t is the Closing Price (CP) of stock in time 't' and CP_{t-1} is the Closing Price (CP) of stock in time't-1'.

7. SentiWordNet

The SentiWordNet is a lexical resource in which every synset of WordNet resource is associated with three numerical scores such as positive, negative and neutral. The semantic information obtained from this resource is incorporated with the technical indicators to forecast the stock price movement.

8. Sentiment Analysis:

The raw data collected from the twitter and money control is preprocessed and sentiment score is to be calculated for every word using (11) and (12):

$$sent_{pos} + sent_{neg} + sent_{neg} = 1 \quad (11)$$

$$sent_{overall} = \sum_{i=1}^{n} \frac{sent_{pos}{}^{i} - sent_{neg}{}^{i}}{n} \quad (12)$$

The overall sentiment and sentiment bullishness is calculated using (13) and (14).

$$\begin{cases} \text{Positive (Pos), if } sent_{overall} \geq +1 \\ \text{Negative } (Neg), \text{ if } sent_{overall} \leq -1 \end{cases} \quad (13)$$

$$Sentiment\ Bullishness\ Index(SBI) = \ln\left[\frac{1+N^{Pos}}{1+N^{Neg}}\right] \quad (14)$$

where N^{Pos} is the total number of positive (Pos) sentiment postings, while N^{Neg} is the total number of negative (Neg) postings. For example, bullish market sentiment is indicated by rising price of stock and falling price of stock indicates the bearish market sentiment. The Sentiment Bullishness Index is

more than 0 is bullish, while 0 is neutral and less than 0 is bearish.

9. Multiple Kernel Learning – Support Vector Machine:

The Multiple Kernel Learning Support Vector Machine (MKL-SVM) integrates the features from more than one sources such as technical, fundamental and sentiment features rather than normal Support Vector Machine (SVM). The kernel used for MKL-SVM is shown in (15).

$$K_{MKL}(x,y) = \sum_{i=1}^{n} \beta_i \, K_i(x,y) \quad (15)$$

with $\beta_i \geq 0, \sum_{i=1}^{n} \beta_i = 1$, where β_i combines weights from sub-kernels $K_i(x,y)$ and MKL is used to estimate the weights for each feature set to find the optimum combined kernels.

D. Prediction and Performance Evaluation:

The Predication and Evaluation component uses MKL-SVM to predict the stock price directional movements for the next trading day. For each feature set uses Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Prediction Accuracy as the evaluation measures.

IV. Experimental Results

A. Correlation Analysis and Granger Causality Test:

The correlation is used to find out the strength of relation between the stock index and sentiment bullishness index. A value close to -1 or +1 indicates a good mathematical fit to a linear model and at the same way the value close to 0 indicates poor fit to a linear model. The value close to +1 denotes a high degree of linear relationship and the value close to -1 denotes a low degree of linear relationship between variables. Although the value = 0 denotes there exists no linear relationship between them. The correlation between trading volume of the each stock and sentiment volume of social media is examined and is shown in Table I. The obtained results show that there is a significant positive correlation for all the stock volume.

TABLE I
Correlation Between Stock Volume and Sentiment Volume

Infosys	HCL	TCS	Tech Mahindra	Wipro
0.831	0.814	0.769	0.741	0.821

The stationarity of each series is to be examined with Augmented Dickey Fuller Unit Root test before applying Granger Causality test [22]. The Granger Causality test is used to examine the social media with the stock price movements is shown in Table II. The obtained results indicate that there is a correlation between social media sentiments and stock market price change. Furthermore, the results show that stock price change is the Granger cause of social media data volume and sentiment polarity (Positive and Negative). But social media data volume is not Granger cause of stock price change while sentiment polarity is the Granger cause of stock price change.

TABLE II
Granger Causality Test Results

Null Hypothesis	Significance Level	Result
Stock Return does not Granger	0.05	Reject
Cause Sentiment Volume	0.01	Accept
Stock Return does not Granger	0.05	Reject
Cause Sentiment Polarity	0.01	Accept
Sentiment Volume does not	0.05	Accept
Granger Cause Stock Return	0.01	Accept
Sentiment Polarity does not	0.05	Reject
Granger Cause Stock Return	0.01	Reject

B. Base Line Methods and Evaluation Measures:

The performance of MKL-SVM is compared with the various baseline methods using RMSE, MAPE, and Prediction Accuracy, are calculated using (16), (17) and (18) respectively.

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(Actual_t - Predicted_t)^2} \quad (16)$$

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{(Actual_t - Predicted_t)}{Actual_t}\right| * 100 \quad (17)$$

$$Prediction\ Accuracy = \left[\frac{No.of\ Right\ Predictions}{Total\ Predictions}\right] * 100 \quad (18)$$

TABLE III
The Result of MAPE

Measures	Infosys	HCL	TCS	Tech Mahindra	Wipro
Technical Indicators (TI)	2.894	4.124	5.243	4.672	3.462
Sentiment Indicators (SI)	3.112	4.762	5.924	5.864	3.962
Fundamental Indicators (FI)	6.243	8.431	9.241	8.981	7.234
All (TI, SI&FI)	1.372	3.934	4.672	4.234	2.632

TABLE IV
The Result of RMSE

Measures	Infosys	HCL	TCS	Tech Mahindra	Wipro
Technical Indicators(TI)	1.431	3.124	3.991	3.921	2.643
Sentiment Indicators(SI)	1.932	3.864	4.296	4.164	2.962
Fundamental Indicators (FI)	2.341	4.624	5.123	4.926	3.641
All(TI, SI&FI)	0.312	1.231	1.892	1.492	0.862

The empirical comparison of MAPE, RMSE and Prediction Accuracy results of five stocks are shown in Tables III, IV and V respectively.

TABLE V
The Result of Prediction Accuracy in %

Measures	Infosys	HCL	TCS	Tech Mahindra	Wipro
Technical Indicators(TI)	81.98	79.12	78.64	79.64	80.24
Sentiment Indicators(SI)	79.69	77.24	76.24	77.12	78.12
Fundamental Indicators (FI)	77.12	76.18	75.74	76.48	76.24
All(TI, SI&FI)	82.98	80.26	79.14	80.16	81.64

Fig. 2 represents the goodness of the proposed system prediction accuracy with the other baseline systems. The obtained results show that accuracy of stock price prediction

accuracy can be significantly improved by the incorporation of semantic enhanced sentiments from the social media. The closing price of stock will go up/down can be predicted more accurately by incorporating the social media sentiments.

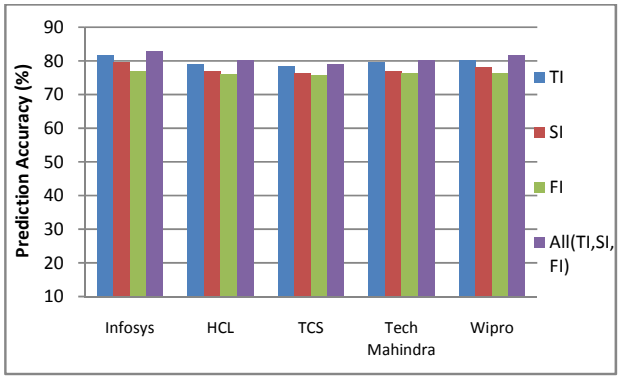

Fig. 2 The Prediction Accuracy of Proposed System with others

V. Conclusion

The stock market forecasting is a very difficult and highly complicated task because it is affected by many factors such as economic conditions, political events and investor's sentiment etc. The stock market series are generally dynamic, nonparametric, noisy and chaotic by nature. The sentiment analysis along with wisdom of crowds can automatically compute the collective intelligence of future performance in many applications. Experimental results show that there is some causal relationship between public sentiment and stock market indices to provide useful investment decisions in the right direction. The accuracy of stock price directional movements can be significantly improved by the incorporation of semantic enhanced sentiments from the social media. The closing price of stock will go up/down can be predicted more accurately by incorporating the social media sentiments. The obtained results indicate that there is a correlation between social media sentiments and stock market price change. Furthermore, the results show that stock price change is the Granger cause of social media data volume and sentiment polarity (Positive and Negative). But social media data volume is not Granger cause of stock price change while sentiment polarity is the Granger cause of stock price change.

References

[1] Shom Prasad Das Sudarsan Padhy, "Support Vector Machines for Prediction of Futures Prices in Indian Stock Market", *International Journal of Computer Applications*, vol. 41 no. 3, 2012.
[2] Tim Loughran and Bill McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks", *The Journal of Finance*, vol.66, no. 1, pp.35–65, 2011.
[3] Johan Bollen, Huina Mao, and Xiao-Jun Zeng, "Twitter mood predicts the stock market*", IEEEComputer*, vol. 44 no.10,pp. 91–94, 2011.
[4] Ling-Chun Hung, "The Presidential Election and the Stock Market in Taiwan". *Journal of Business and Policy Research*, 2011.
[5] Ritanjali Majhi, G., Panda, G., "Prediction of S&P 500 and DJIA Stock Indices using Particle Swarm Optimization Technique", *Proceedings of IEEE Congress on Evolutionary Computation (CEC 2008),* 2008.
[6] Khashei, M., Bijari, M., "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting", *Applied Soft Computing*, vol.11, no.2, pp. 2664–2675,2011.
[7] Chun, C., Qindhua, M., Shuqiang, L., "Research on Support Vector Regression in the Stock Market Forecasting", *Springer – Advances in Intelligent and soft Computing* , vol. 148, pp. 607-612, 2012.
[8] Yung – Keun Kwon and Byung-Ro Moon, "A Hybrid Neruogentic Approach for Stock Forecasting", *IEEE Transactions on Neural Networks*, vol.18, no.3, pp. 851-864, 2007.
[9] Kara, Y., Boyacioglu, M.A., Baykan, O.K, "Predicting Direction of Stock Price Movement Using Artificial Neural Networks and Support Vector Machines: The Sample of the Istanbul Stock Exchange", *Expert Systems with Applications*, vol.38, no.5, pp. 5311 -5319, 2011.
[10] Jeffrey Breen. "Mining twitter for airline consumer sentiment", 2014. (Online; accessed 20-Dec -2014).
[11] Jeff Gentry. Twitter client for r, 2014. (Online; accessed 20-Dec-2014).
[12] Huang, C.F., 2012, "A Hybrid Stock Selection Model Using Genetic Algorithm and Support Vector Machines"., Applied Soft Computing, 12(2), pp. 807-818.
[13] Nuno Oliveira, Paulo Cortez, Nelson Areal, "Some Experiments on Modeling Stock Market Behavior Using Investor Sentiment Analysis and Posting Volume from Twitter", *Proceedings of WIMS'13*, 2013.
[14] Bollen, J., Mao, H., Zeng, X,"Twitter mood predicts the stock markets"., Journal of Computational Science, vol.2, no.1, pp. 1-8, 2011.
[15] Xiangyu Tang, Chunyu Yang, Jie Zhou, "Stock Price Forecasting Combining News Mining and Time Series Analysis", *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology- Workshops*, 2009.
[16] Binoy B. Nair, V.P. Mohandas, N.R. Sakthivel, "A Genetic Algorithm Optimized Decision Tree SVM based Stock Market Trend Prediction System", *International Journal on Compuer Science and Engineering* vol.2, no.9, pp. 2981-2988, 2010.
[17] Nirmala Devi, K., Murali Bhaskaran, V., 2012, "A Semantic Enhanced Approach for Online Hotspot Forums Detection", *Proceedings of IEEE Conference - ICRTIT 2012*,pp. 497-501, 2012.
[18] Vapnik, V, "The Nature of Statistical Learning Theory", Springer-Verlag, pp. 863-884, 2000.
[19] Clerc, M., Kennedy, J, "The particle swarm-explosion, stability, and convergence in a multidimensional complex space", *IEEE Transactions on Evolutionary Computation*, vol.6, no.1, pp. 58-7, 2002.
[20] X. Zhang, H. Fuehres, and P. A. Gloor, "Predicting stock market indicators through twitter i hope it is not as bad as I fear," Anxiety, pp. 1–8, 2009.
[21] Hu M. and Liu B, "Mining and summarizing customer review", *Proceedings of ACM Transactions on Knowledge and Data Engineering*, pp.168-177, 2004.
[22] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods", *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.
[23] http://finance.yahoo.com, http://www.moneycontrol.com