Integrating StockTwits with Sentiment Analysis for better Prediction of Stock Price Movement

Rakhi Batra
Department of Computer Science
Sukkur IBA University
rakhi.bhatra@iba-suk.edu.pk

Sher Muhammad Daudpota
Department of Computer Science
Sukkur IBA University
sher@iba-suk.edu.pk

***Abstract*— Sentiment Analysis is new way of machine learning to extract opinion orientation (positive, negative, neutral) from a text segment written for any product, organization, person or any other entity. Sentiment Analysis can be used to predict the mood of people that have impact on stock prices, therefore it can help in prediction of actual stock movement. In order to exploit the benefits of sentiment analysis in stock market industry we have performed sentiment analysis on tweets related to Apple products, which are extracted from StockTwits (a social networking site) from 2010 to 2017.Along with tweets, we have also used market index data which is extracted from Yahoo Finance for the same period. The sentiment score of a tweet is calculated by sentiment analysis of tweets through SVM. As a result each tweet is categorized as bullish or bearish. Then sentiment score and market data is used to build a SVM model to predict next day's stock movement. Results show that there is positive relation between people opinion and market data and proposed work has an accuracy of 76.65% in stock prediction.**

Keywords—Sentiment Analysis, StockTwits, Stock Prediction, Opinion Mining, Machine Learning, NLP

I. INTRODUCTION

Early research on stock market prediction was created on the Efficient Market Hypothesis (EMH) Fama [1] and the random walk theory [2] [3] [4]. These early models suggested that stock prices cannot be predicted since they are driven by new information (news, blog posts) rather than present/past prices. Thus, stock market prices will follow a random walk and their prediction accuracy cannot exceed 50% [5] which obviously is always there in any random event including tossing a coin. But Baker [6] found that there is effect of sentiment index on securities. They also found that sentiments enhance the predictability of candidate factors of returns.

As an illustration, on April 23, 2013 EST, a tweet from the *Associated Press* (AP) account was posted which stated "Breaking: Two Explosions in the White House and Barack Obama is injured" [7]. The fake tweet, which originated from the hacked AP Twitter account led to an immediate drop in the *Dow Jones Industrial Average* (DJIA). Although the DJIA quickly recovered after an AP disclaimer and a White House press release, this example illustrates the immediate and dramatic effects of perception/news on stock prices. A classical example, indeed.

Sentiment analysis or opinion mining is getting popularity in information technology to determine the sentiments, opinions and emotions. It provides data about comments of people on particular company. Growth of social networking sites contributes in generation of huge quantity of user's comments, opinions and reviews. In order to analyze this information, intelligent systems are being built, that classify the comments and reviews into predefined sentiment classes such as negative, positive or neutral opinion.

The Opinion Mining uses the natural language processing, linguistics computational and text mining to decide whether the review/comment is positive or negative. Text mining or classification is key technique to process the textual data. E-commerce sites, social networking sites, Email filtering systems use the text classification systems to determine the customer's sentiments. Traditionally, models consider the sentiment as binary classification like positive or negative, good or bad, happy or sad. But modern approaches like J48, Naïve Bayes, Support Vector Machine (SVM) and maximum entropy learn from dataset to support multivalued class variable.

SVM is a supervised machine learning algorithm. Mostly it is used in classification but we can also use it in regression. It considers each data point as n-dimensional and plot them in n-dimensional space (n = number of features) where each feature value is value of a particular coordinate. To accurately differentiate the classes, classification is performed by finding hyper-plane. As a result, it learns from classified training data and then outputs an optimal hyper-plane which classifies test data (Figure 1).

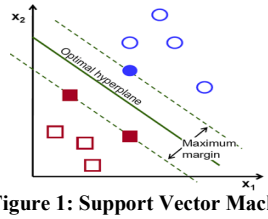

Figure 1: Support Vector Machine

978-1-5386-1370-2/18/$31.00 ©2018 IEEE

By considering the power of opinions on decision making, the idea which we follow in our research is to support the investor's decision in predicting stock movement through sentiment analysis of tweets. For implementation of this idea we will use SVM model of data mining as the sentiment analysis tool. For experimental results we will extract Tweets about Apple from StockTwits from January 2010 to March 2017.

II. Literature Review

Different datasets and techniques have been used by different researchers to predict the stock movement through sentiment analysis. Some of the work is defined in this section.

John et al [8] in their work identify the patterns for confirming correlation between stock prediction and investor sentiments and then predicting the future behavior of stock prices. Naïve Bayes Bernoulli, SVM models were build using Yahoo Finance and Tweeter dataset. 87% accuracy was achieved in prediction with less than 10% error by measuring Accuracy Precision and recall and F-measure of positive negative and neutral. On the other hand Dang et al [9] used the news articles dataset to predict daily stock market direction by performing time series analysis and improved text mining techniques. 73% prediction accuracy was achieved by applying NLTK and linear kernel SVM.

Sun et al [10] used the textual information from user-generated microblogs to predict the stock market. SMF (Sparse matrix factorization) Model was built using S&P 500 and StockTwits data and sentiment were identified with 70.12% accuracy. A different data categorization approach has been proposed in [11]. According to author the average score of feature positivity or negativity cannot be applied to mixed tweets on a single day. Therefore, to increase the quality of feature selection a hybrid SentiWordNet based feature selection approach is proposed. The selected sentimental features are used to train the SVM model.

Zhang et al and Gilbert et al [16] [15] assessed the bloggers sentiment from Live blogpost journal in the polarity of anxiety, fear and worry. They have used Monte Carlo simulation to show the stock movement in S&P 500 index. Sprengers et al [17] evaluated stocks from S&P 100 companies to correlate stock discussions and tweet features that contain Ticker symbols.

Wenga et-al in [12] has collected the data from disparate online data sources like Google, Wikipedia for achieving more accuracy in prediction. Although they have achieved 85% accuracy but no mechanism is defined to validate the data collected from Google and Wikipedia as there can be unauthenticated data. Smailović et al [13] have introduced a new indicator namely positive sentiment probability for finance people to measure the sentiment score of stock. Granger causality test is used to show that sentiment polarity (positive and negative sentiment) can indicate stock price movements a few days in advance.

Rao et al [14] evaluated quantitatively the consequences of twitter sentiments on stock prices movement and to enhance the prediction accuracy results were used with standard model. However, these research works have introduced new dimension to take benefits from public sentiments in development of new successful strategies.

The techniques proposed in literature have provided acceptable outcomes and interesting information about sentiment analysis and the relation between stock market and sentiment analysis. However, the outcomes and results differ and this may depends on choice of sentiment classifier, filtering and preprocessing of tweets and sample of tweets taken for analysis. In our work, we emulate some of techniques from literature and build sentiment analyzer for classification of tweets from StockTwits as bullish or bearish and use this model in conjunction with stock market data to predict the stock movement.

III. Methodology

In this paper, we extracted tweets from StockTwits through pipeline API, processed them for Natural Language Processing (NLP) and sentiment analysis. After that we applied SVM in order to predict the sentiment of each tweet. After predicting sentiment we extracted historical data from Yahoo Finance. We then developed a model for stock market prediction using stock price data and sentiment score to predict the change in stock market.
The proposed methodology for predicting the stock market movement through sentiment analysis is carried out in seven steps (Figure 2).

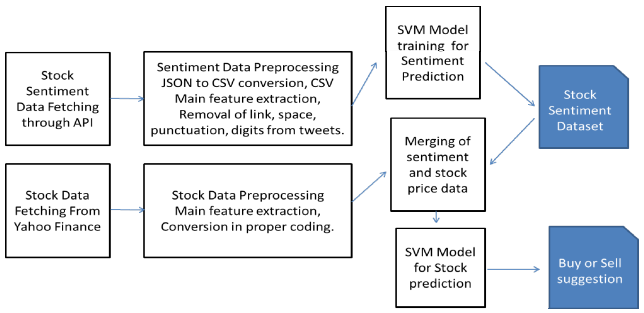

Figure 2: Flowchart of proposed Methodology

1. Stock Sentiment Data Fetching Through API

Sentiment data (Tweets) for Apple has been extracted from StockTwits. It is a social networking platform which provides both large scale text data and high quality data for mining purposes. In 2016, there were around 40 million users of StockTwits worldwide. Efforts are made to filter out financial unrelated tweets and spam messages. It also facilitates user to mark the tweet as bullish or bearish.

It provides two APIs to gather different type of data i-e Search API and Stream API. Streaming API provides the latest 30 tweets of a specific user, company or on provided criteria

978-1-5386-1370-2/18/$31.00 ©2018 IEEE

through query. The user requires authentication key to request the streaming of tweet data. Once the server authenticates user, it opens the connection and stream the tweets to the user.

For our purpose we need historical data of tweet, so, we go for Search API, it requires authentication key. It is RESTful API and provides facility to user for extraction of tweets based on language, region, time, company ticker. Queries have rate limit but we have handled it in our code. The response from server is in the form of JSON objects of tweets. The object contains tweet id, user id, time, tweet text, retweets, sentiment of user on that tweet (bullish and bearish) and more. We have built a utility in C# to retrieve the tweets form 2010 to 2017. Around 300,000 tweets were extracted as JSON object. The JSON data is being converted to CSV file format.

2. Stock Market Data Fetching

Stock market data of Apple is extracted from Yahoo Finance from 2010 to 2017. The data was in excel format and attributes were close price, opening price, low and high price, volume and adjusted close.

3. Sentiment Data Preprocessing

As we know every user has not the same pattern to post a tweet. Users can use numbers, emoticons,, punctuation, special symbols etc. in their posts. For accurate prediction from textual data we preprocessed the tweets to eliminate emoticons, , punctuation, URLs, stop words, numbers etc.

- *Text Processing*

The text of tweet contains words that are inappropriate for sentiment analysis like tweets contains URL, tags, symbols. To remove such words in order to enhance accuracy we have used R commands for handling natural language.

- *Tokenization*

The tweet text is split on basis of spaces to create a list of words, and then those words are used as features to train the classifier.

- *Removing Stopwords*

We excluded the stopwords from list through natural language toolkit (NLTK). It has dictionary of stopwords. Each word in list of words is compared to dictionary words, whenever there is a match, the corresponding word is excluded from list.

- *Twitter Symbols*

Tweets mostly contain symbols like @, #, $, URLs, extra spaces and punctuations for different purposes. All the symbols except $ are removed because they add no value in sentiment analysis.
The words start with $ is ticker of company name so we can filter out them as they may contain useful information for sentiment analysis. To clean up the symbols we used R programing “gsub” function.

After processing, we excluded the unnecessary columns and prepared a CSV file that contains three attributes i-e time, tweet text and sentiment.

4. Stock Data Preprocessing

The stock data extracted from Yahoo Finance was used to decide whether on a particular day the stock price increased or decreased. To make this decision, closing price of today was subtracted from closing price of yesterday, if the result is greater than zero means security (share of a company) price is increased and person can sell the security to earn profit.
On the other side if the difference is less than zero means security price is decreased and person can buy the security or hold if he/she has any. Finally, buy and sell decision was calculated for all days and two columns i.e. date and stock purchase decision was added in data file.

The problem we encountered in data is that as the tweet data is available for all seven days of week but stock data is missing for weekends or whenever market is off. To calculate the missing values we used a function that takes previous and next day value to find the current day’s value. Function is defined below:

$$Y = (PreviousDay + NextDay) / 2$$

5. SVM Model training for Sentiment Prediction

As large amount of tweets data have been collected for training purpose, now we can build and train the classifier. We have used SVM because it is fairly robust to overfitting, it can handle large feature spaces and it is memory efficient. Further, in previous works, SVMs have been shown to be very effective for text categorization.

First the whole data was split into training and test set. To create the training and test set we used the R “createDataFunction” method with 80% training set probability. Then the training set was used to train SVM model and both training and test accuracies were calculated to evaluate the model.

6. Merging of sentiment and stock price data

As we know there could be more than one tweet on each day about one company, so the data that we got from sentiment prediction in previous is aggregated day-wise. It means on a day if there are more positive tweets than negative we say stock sentiment is positive on that day and a person can buy the share.
In this step we merged the stock market data and sentiment data matching date. At the end of pre-processing phase, the file contains three attributes i.e. date, stock price decision and sentiment on a particular day.

978-1-5386-1370-2/18/$31.00 ©2018 IEEE

7. SVM Model for Stock prediction

We used SVM model to suggest whether a person should buy or sell a share. For training the model we used the merged data from step 6. In the data a new attribute of 'decision' was added that is prediction variable. The value of this attribute for training set was calculated as follows: if the stock and sentiment both are positive the decision is buy otherwise sell.
Whole data was split into training and test set and then model was trained with training set and evaluated by calculating prediction accuracy of test set.

IV. RESULTS

The models that we built during implementation phase were evaluated by calculating training and test accuracy, Recall and Precision. Results are presented in the form of confusion matrix (Table 1).

	True	False
True	TP	TN
False	FP	FN

Table 1: Confusion matrix for TP, TN, FP, FN

Where TP (True Positive), FP (False Positive) TN (True Negative), and FN (False Negative) can be defined as follows:
True Positive (TP): Number of positive tuples which are correctly classified by classifier.
True Negative (TN): Number of negative tuples which are correctly classified by classifier.
False Positive (FP): Number of negative tuples which are classified as positive by classifier.
False Negative (FN): Number of positive tuples which are classified as negative by classifier.

***Measurement Factors**:*

Accuracy: measure the closeness of calculated value to the known or standard value.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

Precision: measure the closeness of two or more measured values to each other.

$$\Pr ecision = \frac{tp}{tp + fp}$$

Recall: is the fraction of positive instances that have been retrieved over the total number of positive instances. It has also called sensitivity.

$$\mathrm{Re}\,call = \frac{tp}{tp + fn}$$

As we have created two SVM models i-e one for sentiment analysis of Stock tweets and model for predicting stock movement. The results are presented for both models. We achieved 91.2% training accuracy in SVM Sentiment model with 98% recall, and 90.9% Precision. For SVM Sentiment model the achieved test accuracy is 63.5%, 75.3% Recall, and 76.8% Precision. The training and test results are presented in contingency matrix in Table 2 and Table 3 respectively.

Actual/Predicted	True	False
True	31240	613
False	3091	7490

Table 2: Sentiment Analysis Training Confusion Matrix

Actual/Predicted	True	False
True	4051	1325
False	1222	385

Table 3: Sentiment Analysis Testing Confusion Matrix

We achieved 75.22% training accuracy in SVM Stock Model with 100% Recall, and 66.7% Precision. For SVM Stock Model the achieved test accuracy is 76.68%, 100% Recall, and 69.5% Precision. The training and test results are presented in contingency matrix in Table 4 and

Table 5 respectively.

Actual/Predicted	True	False
True	544	0
False	271	279

Table 4: Stock Prediction Training Confusion Matrix

Actual/Predicted	True	False
True	389	0
False	170	170

Table 5: Stock Prediction Testing Confusion Matrix

V. CONCLUSION

Early research on Stock Market prediction were totally based on random walks and numerical prediction but with the introduction of behavioral finance, the people's belief and mood were also considered while predicted about stock movement. Making it more efficient we used the idea of sentiment analysis of Stock Tweets through machine learning models.
We implemented the idea by collecting sentiment data and stock price market data and built an SVM models for prediction and in the last we measured the prediction accuracy. Results showed that we have achieved 75.22% training

978-1-5386-1370-2/18/$31.00 ©2018 IEEE

accuracy and 76.68% test accuracy. It can be improved if we increase the size of data set.

VI. Limitations and Future work

Limitation:

As the StockTwits data was downloaded by using API and attributes were selected on need basis. Any well-known dataset, reviewed by researchers can be used to improve the authenticity of results.

Future Work:

There is need to improve the test accuracy of text classification algorithm. More training data can be used to enhance the prediction accuracy. This can be extended to support decision making for more investment options like commodity market and real state.

VII. References

[1] E. F. Fama, "The Behavior of Stock-Market Prices," *The Journal of Business,* pp. 34-105, 1965.

[2] P. H. Cootner, "The Random Character of Stock Market," *The Journal of Business.*

[3] E. F. Fama, " Random walks in stock market prices.," *Financial Analysts Journal.*

[4] E. F. Fama, L. Fisher, M. C. Jensen and R. Roll, " The adjustment of stock prices to new information.," *International Economic Review.*

[5] J. Bollen, H. Mao and X. Zeng, " Twitter mood predicts the stock market.," *Journal of Computational Science,* 2011.

[6] J. W. Malcolm Baker, "Investor Sentiment and the Cross-Section of Stock Returns," 2006.

[7] F. M. Megahed and Jones-Farmer, in *Frontiers in statistical quality control 11* ., 2015.

[8] J. Kordonis, "Stock Price Forecasting via Sentiment Analysis on Twitter," in *The 20th Panhellenic Conference on Informatics (PCI '16)*, Greece, 2016.

[9] M. Dang, "Improvement Methods for stock market prediction using financial news articles," in *Information and Computer Science (NICS), 2016 3rd National Foundation for Science and Technology Development Conference*, Danang, Vietnam, 2016.

[10] A. Sun, "Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction," *International Review of Financial Analysis,* pp. 272-281, 2016.

[11] P. Meesad, "Stock trend prediction relying on text mining and sentiment analysis with tweets," in *Information and Communication Technologies (WICT), 2014*, Bandar Hilir, Malaysia, 2014.

[12] B. Wenga, M. A.Ahmed and F. M.Megahedbc, "Stock market one-day ahead movement prediction using disparate data sources," *Expert Systems with Applications,* pp. 153-163, 2017.

[13] J. Smailović, M. Grčar, N. Lavrač and M. Žnidaršič, "Predictive Sentiment Analysis of Tweets: A Stock Market Application," in *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* , Verlag Berlin Heidelberg , Springer, 2013, pp. 77-88.

[14] T. Rao and S. Srivastava, "Analyzing Stock Market Movements Using Twitter Sentiment Analysis," in *International Conference on Advances in Social Networks Analysis and Mining.*, 2012.

[15] E. Gilbert and K. Karahalios, "Widespread worry and the stock market," *Artifical Intelligence,* 2010.

[16] X. Zhang, H. Fuehres and P. A. Gloor, "Predicting stock market indicators through twitter," 2009.

978-1-5386-1370-2/18/$31.00 ©2018 IEEE