Predicting Stock Movement using Sentiment Analysis of Twitter Feed

Pranjal Chakraborty, Ummay Sani Pria, Md. Rashad Al Hasan Rony, Mahbub Alam Majumdar
Department of Computer Science and Engineering, BRAC University
66, Mohakhali, Dhaka, Bangladesh
[1]pcborty@outlook.com
[2]ummaysani@gmail.com
[3]rah.rony@gmail.com
[4]majumdar@bracu.ac.bd

***Abstract*—Collecting data from social networking sites is a popular way of opinion mining. These opinions show the sentimental state of a large number of people. In this paper, we have shown how much we can predict stock movement from twitter's tweets sentiment analysis. Our work is done on one year's (2016) data of tweets that contained 'stock market', 'stocktwits', 'AAPL' keywords. 'AAPL' related tweets were used to see if these tweets can predict the company's stock indices whereas 'stock market', 'stocktwits' related tweets for predicting the stock market movement of US. Since we are predicting the stock values, we used Boosted Regression Tree model for this purpose.**

Keywords— Tweets; Sentiment analysis; Decision tree; Opinion mining; Machine Learning; Random forest; Decision tree; Boosted tree; and Support Vector Machine (SVM)

I. INTRODUCTION

Social networking sites now have become an integral part of human life. Twitter, which also features the tag "microblogging site", is such a social networking site in action since 2006. 100 million active Twitter users update nearly 500 million tweets [12] every day. Users express their opinion, decisions, feeling etc. through these tweets, which can be translated into useful information.

Previously, we have seen sentiment analysis has been done on IMDB movie reviews, even on Twitter feed. Most of the recent applications of sentiment analysis lies more upon kernel based classifiers.

In this paper, we are going to use a sentiment tagged Twitter dataset of 1.5 million tweets, collected from Sentiment140 [10] for sentiment classification. We have filtered the tweets by some regularly used and some unique parameters to convert them as close as possible to plain text. Then, we used Boosted Regression Tree classifier for predicting next day's stock movement with present day's tweets containing 'stock market', 'stocktwits', 'AAPL'.

II. LITERATURE SURVEY

Sentiment relates to feelings, attitudes, emotions and opinions. Social media sentiment analysis is an excellent source of information and can provide insights that can determine marketing strategy, improve campaign success and many more. In past years, there has been a fair amount of research on sentiment analysis on movie reviews, twitter feeds.

Agarwal, Xie, Vovsha, Rambow, and Passonneau examined sentiment analysis on Twitter Data and introduced POS-specific prior polarity features and explored the use of a tree kernel to remove the need for tedious feature engineering. Their introduced features and tree kernel performed approximately at the same level, both outperforming the state-of-art baseline [1].

Pang and Lee proposed a novel machine-learning method that applies text-categorization techniques to just the subjective portions of the document. Extracting these portions can be implemented using efficient techniques for finding minimum cuts in graphs. They used mincut to improve the classification of a sentence into either 'objective' or 'subjective', with the assumption that sentences close to each other tend to have the same class. Convolutional Neural Networks (CNN) can capitalize on distributed representations of words by first converting the tokens comprising each sentence into a vector, forming a matrix to be used as input [2].

Kim proposed a simple one-layer CNN that achieved state-of-the-art results across several data sets. The very strong results achieved with this comparatively simple CNN architecture suggest that it may serve as a drop-in replacement for well-established baseline models, such as Support Vector Machine (SVM) ([4] Joachims, 1998) or Logistic Regression. Kim defined a one-layer CNN architecture that uses pre-trained word vectors as inputs, which may be treated as task-specific or static vectors. In that approach, word vectors are treated as

978-1-5386-1023-7/17/$31.00 ©2017 IEEE

fixed inputs, while in the latter they are 'tuned' for a specific task. Learning task-specific vectors through fine tuning gave them better result [3].

Efthymios Kouloumpis, Theresa Wilson, Johanna Moore have used Hashtag and Emoticon in tweets to predict sentiment using n-gram, lexicon, parts-of-speech, micro-blogging as features [5].

Peter D. Turney worked on an unsupervised classification algorithm called PMI_IR (Turney 2001), using the association of adjectives in reviews [6].

Sida Wang and Christopher D. Manning observed that the performance of Naïve Bayes, SVM varies with bigram feature, length of documents without using stopwords, lexicon [7].

III. DATA

Our dataset contains, as mentioned earlier, 1.5 million (1, 578, 614) hand-tagged tweets, collected through Sentiment140 API. The tweets are tagged '1' and '0' for being 'positive' and 'negative' respectively. Then we performed a random split over the dataset, to divide the dataset into training dataset, containing 1, 499, 337 tweets and testing dataset, containing 79, 277 tweets (table 1).

We collected tweets containing 'stock market', 'stocktwits', 'AAPL' keywords from January to December 2016 for predicting stock movement. For each day, we downloaded at most a thousand tweets containing those keywords.

Also, we used DJIA (Dow Jones Industrial Average), APPLE Inc. closing stock difference values from January to December 2016. We took the closing prices as it shows, after all the ups and downs how the stock point remains at the end of a day.

As, the stock keywords related huge data was not hand tagged, we used the Sentiment140 data to get the sentiment values of those tweets. The distribution of positive and negative tweets of Sentiment140 is almost equal in both training and testing data.

Table 1. Data distributions

	Training	Testing
Positive	750, 527	39, 651
Negative	748, 810	39, 626

Table 2. First 8 rows of raw data

Sentiment	Sentiment Text
0	is so sad for my APL friend.....
0	I missed the New Moon trailer...
1	omg its already 7:30 :O
0	..Omgaga. Im soooim gunna CRy. I've been at this dentist since 11..I was suposed 2 just get a crown put on (30mins)...
0	i think mi bf is cheating on me!!!T_T
0	or i just worry too much?
1	Juuuuuuuuuuuuuuuussssst Chillin!!

IV. DATA PREPROCESSING

The tweets have gone through following pre-processing steps-

(i) All characters have been changed to lower case letters to avoid repetition of same words in feature vector.

(ii) All URL, starting with <http://>, <https://> and <www. > have been stripped off from the text.

(iii) Consecutive whitespaces are replaced by single whitespace. Apart from that, more than two consecutive characters, symbols, punctuation marks have been replaced by just two of them.

(iv) Usernames (starts with <@>) are removed. Hashtags are kept stripping off the <#> sign. Emoticons are kept the same.

(v) Apart from these, all retweet signs ('RT'), single (<'>) and double (<">) quote signs are stripped off. All non-ascii characters are removed too.

After preprocessing, data of table 2 took the form of table 3.

Table 3. First 8 rows of filtered data

Sentiment	Sentiment Text
0	is so sad for my apl friend..
0	i missed the new moon trailer..
1	omg its already 7:30 :o
0	..omgaga. im soo im gunna cry. i've been at this dentist since 11..i was suposed 2 just get a crown put on (30mins)..

0	i think mi bf is cheating on me! t_t
0	or i just worry too much?
1	juusst chillin!

V. RESULTS OF SENTIMENT ANALYSIS

GraphLab Create has been used to work with this large amount of data. 5% of the training data has been used for validation. These results in table 4, came from Sentimet140 dataset, by applying different classification algorithms along with unigram, bigram, trigram approach.

Table 4. n-gram (n=1, 2, 3) model outputs

Model	Unigram		Bigram		Trigram	
	Train	Test	Train	Test	Train	Test
Logistic Regression	0.84	0.78	0.98	0.76	0.98	0.74
SVM	0.84	0.79	0.98	0.76	0.98	0.74
Decision Tree	0.64	0.64	0.52	0.52	0.52	0.52
Boosted Tree	0.55	0.55	0.60	0.60	0.55	0.54
Random Forest	0.64	0.64	0.53	0.53	0.53	0.53

Here from table 4 we can see that, SVM model gave the best accuracy with unigram approach on the test tweets. Although, SVM performs better, logistic classifier performed almost as well as SVM.

Table 5. Complete information of test data evaluation

	Accuracy	F1 score	Precision	Recall
Logistic Reg.	0.7886	0.7879	0.7909	0.7849
SVM	0.7908	0.7912	0.7923	0.7901
Decision Tree	0.6415	0.6649	0.6244	0.7112
Boosted Tree	0.684	0.6962	0.6705	0.7239
Random Forest	0.6428	0.6688	0.6236	0.7211

In table 5, The precision is the ratio t_p / (t_p + f_p) where t_p is the number of true positives and f_p is the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. The recall is the ratio t_p / (t_p + f_n) where t_p is the number of true positives and f_n the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0. The F-beta score weights recall more than precision by a factor of beta; beta = 1. 0 means recall and precision are equally important.

VI. STOCK MOVEMENT PREDICTION

From the results of sentiment analysis, we found that SVM worked best on our test data of Sentiment140, along with unigram feature. So, for sentiment analysis of the stock related tweets, we used SVM with unigram approach where Sentiment140 data set is our training data.

Tweets containing 'stock market', 'stocktwits' were trained to predict general stock market points (DJIA) closing difference. Besides, tweets containing 'AAPL' were trained to predict Apple Inc. closing stock point difference. Tweets were ignored of the days when the stock market or Apple's stock-prices was closed. All of these tweets went through the data formatting mentioned in the paper.

VI A: SCORE OF 'STOCK MARKET', 'STOCKTWITS', 'AAPL' RELATED TWEETS

In our work, we have used marginal values that came as an output from running Support Vector Machine (SVM) on tweets containing 'stock market', 'stocktwits', 'AAPL'. As SVM is basically a classifying algorithm that draws the best hyperplane or margin to separate classes, marginal values are distance from the hyperplane to a data point.

We took the average marginal value of these tweets for each day as we had at most a thousand tweets for each day. Positive marginal value refers to positive sentiment and negative marginal value refers to negative sentiment.

VI B: STOCK INDEX VALUE PREDICTION USING BOOSTED REGRESSION TREE

We used the Boosted Regression Tree model for the stock index value prediction. Training set is data from January to August 2016 and testing was done on stock related data from September to December

2016. The marginal values of tweets containing 'stock market', 'stocktwits' are trained with DJIA closing price difference. Whereas, tweets that contain 'AAPL'; their marginal value is trained with APPLE Inc. closing stock price difference.

We trained our model to predict stock price difference of the next day. In the training data set average marginal value are of a day's tweets and their corresponding closing price difference is between that day and the next day. So that, after getting marginal value of tweets of present day, we can predict how much stock market will rise or fall the next day. In other words, for present day's stock value prediction we will need previous day's tweets average marginal value.

The tweets from September to December 2016 which were used for testing went through SVM classification first, to obtain average marginal values as our Boosted Regression Tree model is trained with average marginal values. These average marginal values were then used to predict next day's stock difference by our Boosted Regression Tree model.

Table 6. First few entries to train Boosted Regression Tree model of the tweets containing 'stocktwits'

Date	Average Marginal Value of tweets	Actual Closing price difference
04/01/2016	-5.1195209	9.72
05/01/2016	13.2967218	-252.15
06/01/2016	22.6936427	-392.41

In the table 6, the average marginal value of tweets of 04/01/2016 is -5.1195209 and the next day closing price increased by 9.72 points. So, if given today's tweets based on our training set we can predict tomorrow's stock price difference in advance.

VII: PREDICTION RESULTS OF STOCK MOVEMENT

We have plotted both actual stock difference and predicted stock difference of testing period (September to December 2016).

For fig.1, fig.2, and fig.3 time is on the x-axis and difference of stock points are on the y-axis. Whereas, for fig.4, and fig.5, time is on the x-axis and stock points are on the y-axis. Here, fig.4, and fig.5 are based on the calculated stock indices using predicted stock points differences.

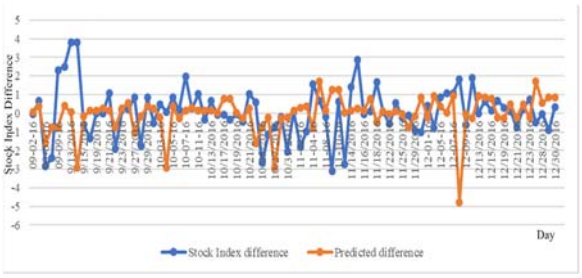

Fig.1: Company specific ('AAPL') closing stock price difference prediction using tweets containing 'AAPL' with Boosted Regression Tree.

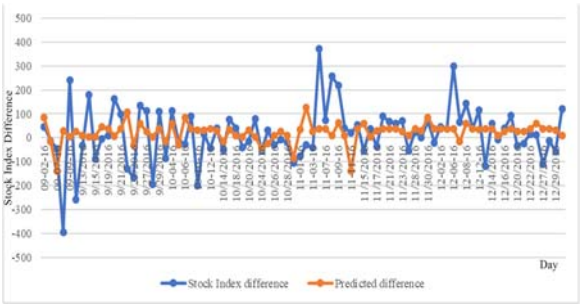

Fig.2: Prediction of Closing Stock Price difference value using (Boosted Regression Tree) on tweets containing 'stock market'.

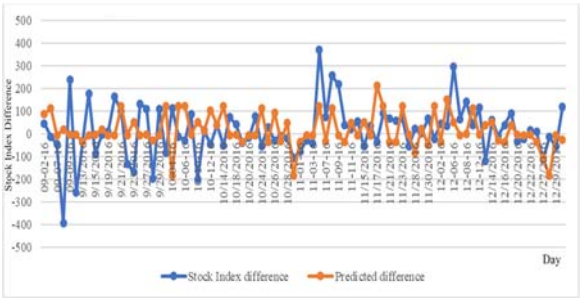

Fig.3: Prediction of Closing Stock Price difference value using (Boosted Regression Tree) on tweets containing 'stocktwits'.

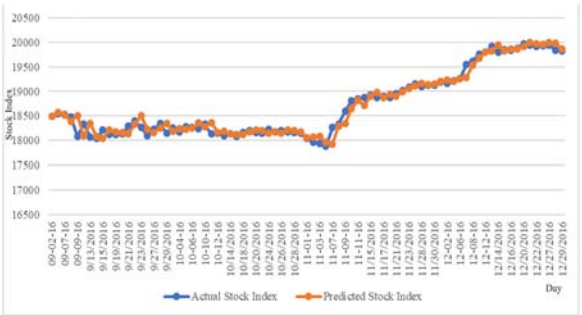

Fig.4: Actual closing stock index vs. Predicted closing stock index graph of the tweets containing 'stock market'

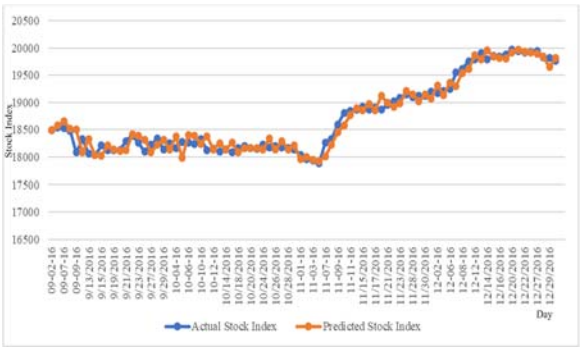

Fig.5: Actual closing stock index vs. Predicted closing stock index graph of the tweets containing 'stocktwits'

Table 7. Error metrics

Tweets with phrase	**Mean Absolute Error of Predicting Closing Stock price difference**	**RMSE**
'stock market'	82.95103119	116.1048303
'stocktwits'	103.3456242	131.3553
'AAPL'	1.201669284	1.731554777

In the table 7, Root Mean Square Error (RMSE) value is calculated using the following equation, where y_t is the actual closing stock difference and $\hat{y}_t$ is the predicted one.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}$$

Besides, Mean Absolute Error is calculated using the following equation, where y_t is the actual closing stock difference and $\hat{y}_t$ is the predicted one.

$$Mean\ absolute\ \ error = \frac{\sum_{t=1}^{n}|\hat{y}_t - y_t|}{n}$$

From the plotted graphs of closing price difference (fig.1, fig.2, fig.3), we can see that, in the first few days of September, stock points had major jumps in rise and fall. As these high rise and falls are quite unexpected and could not be predicted well by our model, for those few days the total error increased. Otherwise from the graphs we can see that our model predicted the values quite well.

The values in fig.4, and fig.5 are calculated using following equation, where, y_t is the predicted stock index of the next day and y_{t-1} is the actual stock index of the present day. Δy is the predicted stock difference of the next day.

$$y_t = y_{t-1} + \Delta y$$

VIII. CONCLUSION AND FUTURE WORK

In our work, we tried to predict the future movement of the stock market of United States by analyzing sentiment of Twitter posts, which are related to Stock market. In order to do this, we collected stock related tweets, obtained their average marginal value by using SVM. After that, we prepared the training set with those tweets and with corresponding DJIA or Apple Inc. closing stock index difference between present day and next day. Then we tested on similar stock related tweets on a different timeline to see how much we can predict stock index.

Tweets are generally less informative, misspelled and often grammatically incorrect, which makes it harder to classify with traditional classifiers. Though we found that SVM performs better for classifying Twitter feeds for sentiment analysis, our future work would be to implement a sentiment classifier with Recurrent Neural Network (RNN) and to predict the movement of different Stock markets.

From our result, it is clear that, extremely high and extremely low difference in Stock Indexes is difficult to predict with Boosted Regression Tree. However, except for those days, our model predicted very well on the given data set. Since our work is done on data of one year, performance may be improved by training on larger data set.

IX. REFERENCES

[1] Agarwal, A., Xie, B.,Vovsha, I.,Rambow, O.,& Passonneau, R. (2011). Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media, LSM '11, pp.30–38, Portland, Oregon.

[2] Pang and L. Lee.2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In ACL-2004.

[3] Y. Kim, "Title: Convolutional neural networks for sentence classification, " 2014. [Online]. Available: https://arxiv. org/abs/1408.5882.

[4] T. Joachims, "Text categorization with support vector machines: Learning, " 1998.

[5] E. Kouloumpis, T. Wilson, J. Moore, “Twitter Sentiment Analysis: The Good the Bad and the OMG!” 2011.

[6] P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, " 2002.

[7] S. Wang, C. D. Manning, “Baselines and Bigrams: Simple, Good Sentiment and Topic Classification,” 2012.

[8] Y. I. Chang, "Boosting SVM Classifiers with Logistic Regression, ".

[9] A. Kennedy and D. Inkpen, "SENTIMENT CLASSIFICATION of MOVIE REVIEWS USING CONTEXTUAL VALENCE SHIFTERS, " Computational Intelligence, vol.22, no.2, pp.110–125, May 2006.

[10] "API - sentiment140 - A Twitter sentiment analysis tool, ". [Online]. Available: http://help. sentiment140. com/api. Accessed: Dec.20, 2016.

[11] K. Inc, "Datasets, " 2017. [Online]. Available: https://www. kaggle. com/datasets. Accessed: Dec.12, 2016.

[12] "Twitter basics, " in Twitter, 2017. [Online]. Available: https://business. twitter. com/en/basics. html. Accessed: Feb.1, 2017.

[13] P. Temin, "The Great Recession and the Great Depression", National Bureau of Economic Research, vol.15645, 2010.

[14] B. Malkiel, "Efficient Market Hypothesis", The New Palgrave: Finance. Norton, New York, pp.127-134, 1989.

[15] S. Lai, L. Xu, K. Liu and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification", Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[16] A. Kennedy and D. Inkpen, "Sentiment Classification of Movie Reviews Using Contextual Valence Shifters, " Computational Intelligence, vol.22, no.2, pp.110–125, May 2006.

[17] P. Diamond, “Behavioral Economics,” Massachusetts Institute of Technology, Dept. of Economics, 2008.

[18] C. Bishop, “Pattern Recognition and Machine Learning,” Springer, 2006.