IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 16-20
www.iosrjournals.org

Prediction Algorithm using Lexicons and Heuristics based Sentiment Analysis

Aakash Kamble[1], Darshan Vakharia[2], Rachit Verma[3], Dr. Rajesh Bansode[4]
[1, 2, 3]*(Information technology, Thakur College of Engineering & Technology, Mumbai, India)*
[4]*(Associate Professor, Thakur College of Engineering & Technology, Mumbai, India)*

Abstract : *Stock market prediction has been a vital requirement of the investors. Computer science plays an important role in it. Well-organized and EMH had been one of the prominent theory about stock prediction. Collapse of it had resulted in research in the area of prediction of stocks. The idea is taking non quantifiable data such as financial tweets related to a company and predicting its stock with tweets (twitter) sentiment classification. Considering the fact that twitter data have impact on stock market, this is an attempt to figure out relationship between twitter data and stock trend. Classification models are created which depict polarity of tweets being positive or negative. Proposed model is assessed for various data and promising results are obtained. Sentimental analysis based on lexicons and heuristics provided solidity*

Keywords – *Stock Prediction, Sentiment Analysis, Lexicons, Heuristics, Prediction Algorithm*

I. Introduction

Stock market and its trends have extremely volatile nature. It attracts researchers to understand volatility and predict next moves of market. Investors and market analysts study the market behavior and plan their investment strategies accordingly. As tremendous amount of data is generated by stocks every day, it is a difficult job for an individual, analyzing all the past and present information for predicting future trend of a stocks. Technical analysis and other is Fundamental analysis are two major methods of predicting stock market trend. Technical analysis considers past price and volume to predict the future trend whereas Fundamental analysis. On the other hand, Fundamental analysis of a business involves analyzing its financial data to get some insights. The efficacy of both are always a topic of dispute by the efficient-market hypothesis according to which the stock market prices are extremely unpredictable.

II. Literature Survey

Stock price trend prediction is an active area of research, as more accurate predictions are directly proportional to more returns in stocks. Which is evident by the fact that in recent years, significant efforts have been put into developing models to effectively predict future trend of a specific stock or entire market. Most of the techniques which are existing make use of technical indicators. Some of the researchers showed that there is a strong relationship between news article about a company and its stock prices fluctuations [2]. Our proposed idea is supported by Bollen et al’s [1] theory. Following is discussion on previous research on sentiment analysis of text data and different classification techniques.

Nagar and Hassler in their research [3] presented an automated text mining based approach to aggregate news stories from various sources and create a News Corpus. The Corpus is filtered down to relevant sentences and analyzed using Natural Language Processing (NLP) techniques. A sentiment metric, called Twitter Sentiment, utilizing the count of polarity (negative or positive) words is proposed as a measure of the sentiment of the overall news corpus. Various open source packages and tools are used by them to develop the news collection and aggregation engine as well as the sentiment evaluation engine. They also state that the time variation of Twitter Sentiment shows a very strong correlation with the actual stock price movement. Yu et al [4] present a text mining based framework to determine the sentiment of news articles and illustrate its impact on energy demand. News sentiment is quantified and then presented as a time series and compared with fluctuations in energy demand and prices. J. Bean [5] uses keyword tagging on Twitter feeds about airlines satisfaction to score them for polarity and sentiment. This can provide a quick idea of the sentiment prevailing about airlines and their customer satisfaction ratings. We have used the sentiment detection algorithm based on this research.

This research paper [6] studies how the results of financial forecasting can be improved when Twitter data with different levels of relevance to the target stock are used simultaneously. They used multiple kernels learning technique for partitioning the information which is extracted from different five categories of news articles based on sectors, sub-sectors, industries etc. Twitter data are divided into the five categories of relevance to a targeted stock, its sub industry, industry, group industry and sector while separate kernels are employed to analyze each one. The experimental results show that the cumulative use of various tweets categories increases the prediction performance in comparison with methods based on a lower number of news categories. It shows that highest prediction accuracy and return per trade were achieved for MKL when all five categories of tweets were utilized with two separate kernels of the polynomial and Gaussian types used for each tweets category.

III. Related Work

Our work is based on Bollen et al’s strategy [1] which received widespread media coverage recently. They also attempted to predict the behavior of the stock market by measuring the mood of people on Twitter. The authors considered the tweet data of all twitter users in 2008 and used the Opinion Finder and Google Profile of Mood States (GPOMS) algorithm to classify public sentiment into 6 categories, namely, Calm, Alert, Sure, Vital, Kind and Happy. They cross validated the resulting mood time series by comparing its ability to detect the public’s response to the presidential elections and Thanksgiving Day in 2008. They also used causality analysis to investigate the hypothesis that public mood states, as measured by the Opinion Finder and GPOMS mood time series[7], are predictive of changes in DJIA closing values. Self Organizing Fuzzy Neural Networks is used by the researchers to predict DJIA values using previous values. A remarkable accuracy of nearly 87% in predicting the up and down changes in the closing values of Dow Jones Industrial Index (DJIA) is shown by their results [3].

IV. Proposed Work

The Efficient Market Hypothesis (EMH) states that stock market prices are largely driven by new information and follow a random walk pattern [8]. Though this hypothesis is widely accepted by the research community as a central paradigm governing the markets in general, several people have attempted to extract patterns in the way stock markets behave and respond to external stimuli. These moods and previous days’ Dow Jones Industrial Average (DJIA) values are used to predict future stock movements and then use the predicted values in our portfolio management strategy.

Market price of the equity shares is very difficult to predict and it is based on historical prices of the stock but it is an important tool for short term investors for achieving maximum profits. MLP neural networks have been the existing method for price predictions. However, MLP neural network does not always give accurate prediction in case of volatile markets. In proposal we are introducing a new way of collaborating both neural networks and decision tree to forecast the stock market price more accurately than MLP.

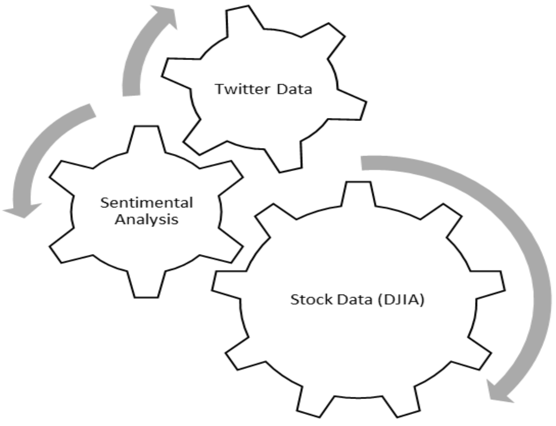

Fig. 1: Core mechanism

This design can logically be seen as harmonic working of three blocks, first is the twitter data, second is sentimental analysis model and third is the stock data in form of DJIA. The twitter data in form of raw text from tweets is collected feed into sentimental analysis model. The output is twitter data with its’ polarity score. The relationship between the stock and data with polarity score is obtained. The stock data and the twitter data is plotted and the predicted output is obtained [1].

Moods and previous days’ Dow Jones Industrial Average (DJIA) values are used to predict future stock movements and then use the predicted values in portfolio management strategy results show a remarkable accuracy in predicting the up and down changes in the closing values of Dow Jones Industrial Index (DJIA). The paper introduces a way of combining decision and tree neural networks to predict the stock market price better than MLP.

V. Methodology

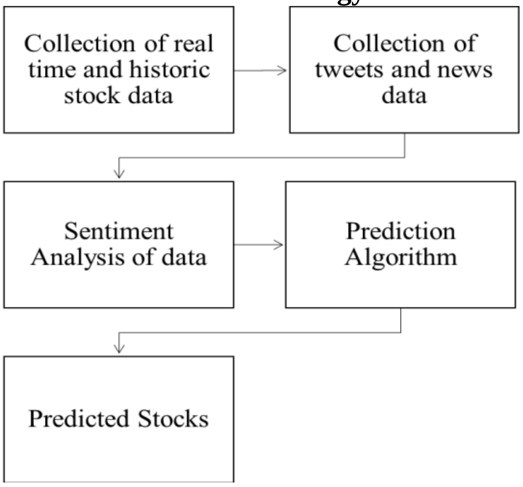
Collection of real time and historic stock data
Collection of tweets and news data
Sentiment Analysis of data
Prediction Algorithm
Predicted Stocks

Fig. 2: Process Flow

1. Sentiment Analysis Model:

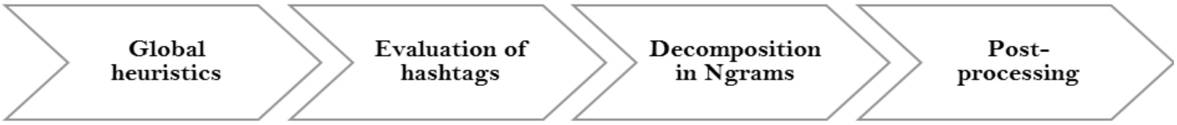

Fig. 3: Classification Engine of proposed Sentiment Analysis Model

The four steps of the text (Twitter tweets) classification.

1.1 Global heuristics:

Smileys and onomatopes carry strong indications of sentiment, but also come in a variety of orthographic forms which require methods devoted to their treatment. Whatever the negative sentiments (Hate you) signaled in the tweet, the final smiley has an overriding effect and signals the strongest sentiment in the tweet. For this reason smileys located in final positions are recorded as such [9].

1.2 Evaluation of hashtags:

Hashtags are of special interest as they single out a semantic unit of special significance in the tweet. Exploiting the semantics in a hashtag faces the issue that a hashtag can conflate several terms, as in #greatstuff or #notveryexciting. Application of a series of heuristics matching parts of the hashtag with lexicons. In the case of #notveryexciting , the starting letters not will be identified as one of the terms in the lexicon for negative terms. Similarly, the letters very will be identified as one of the terms present in the lexicon for “strength of sentiment” exciting will be detected as one of the terms in the lexicon for positive sentiment. Taken together, not very exciting will lead to an evaluation of a negative sentiment for this hashtag. This evaluation is recorded and will be combined with the evaluation of other features of the tweet at a later stage [9].

1.3 Decomposition in Ngrams:

The text of the tweet is decomposed in a list of unigrams, bigrams, trigrams and quadrigrams. For example, the tweet This service leaves to be desired will be decomposed in list of the following expressions: “This, service, leaves, to, be, desired, This service, service leaves, leaves to, to be, be desired, This service leaves, service leaves to, leaves to be, to be desired, This service leaves to, service leaves to be, leaves to be desired” The reason for this decomposition is that some markers of sentiment are contained in expressions made of several terms. In the example above, to be desired is a marker of negative judgment recorded as such in the lexicon for negative sentiment, while desired is a marker of positive sentiment. The model loops through all the n-grams of the tweet and checks for their presence in several lexicons. If an n-gram is indeed found to be listed in one of the lexicons, the heuristic attached to this term in this lexicon is executed, returning a classification (positive sentiment, negative sentiment, or another semantic feature) [9].

1.4 Post-processing:

At this stage, the methods described above may have returned a large number of (possibly conflicting) sentiment categories for a single tweet. For instance, in the example.
"This service leaves to be desired, the examination of the n-grams has returned a positive sentiment classification (desired) and also negative (to be desired)."
A series of heuristics adjucates which of the conflicting indications for sentiments should be retained in the end. In the case above, the co-presence of negative and positive sentiments without any further indication is resolved as the tweet being of a negative sentiment. If the presence of a moderator is detected in the tweet (such as but, even if, though), rules of a more complex nature are applied [9].

2. Proposed Prediction Algorithm:

```
inputs : stockItem, stockSentiment
local variables: predictedPrice, priceChange, var, count
priceChange <- 0
count <- 0

while : count < 5
var <- stockSentiment.getPositive() – stockSentiment.getNegative()
        priceChange <- priceChange + (var * 100) / stockSentiment.getTotalTweets()
        count ++
end

priceChange <- priceChange/5;
predictedPrice <- stockitem.currentPrice + priceChange;

return predictedPrice;
```

VI. Results and Discussions

Sr. No.	Entity	Number of tweets	Predicted stock Price ($)	Actual stock Price - next day ($)	Deviation	% Deviation
1	Alphabet Inc.	675	809.44	815.91	6.47	0.795
2	Facebook Inc.	900	128.42	128.74	0.32	0.249
3	Yahoo Inc.	900	42.08	42.26	0.18	0.02
4	Apple Inc.	898	120.22	120.03	0.19	0.158

Table 1: Comparing predicted price with actual price.

The comparison is done among four companies (Alphabet Inc., Facebook Inc., Yahoo Inc., Apple Inc.). Prediction of stock prices is done using the proposed algorithm. The reading of Actual price is taken next day. The average deviation is 1.859 and mean present deviation is 0.305%. It has been observed that in case of analysis data with low number of tweets the deviation is found to be more than that observed otherwise. With increase in the dataset and available of more structured stock data the accuracy will increase and proportionally the deviation will decrease accordingly.

SR .NO.	ENTITY	SENTIMENT	PRICE (Actual)
1	Alphabet Inc.	Positive	Increase
2	Facebook Inc.	Positive	Increase
3	Yahoo Inc.	Negative	Decrease
4	Apple Inc.	Positive	Increase

Table 2: Mapping of Sentiment to Actual price

The correlation between the sentiment and the change in price is found to be high. With the accuracy being 100% for the considered entities.

VII. Future Scope

We would like to extend this research by adding more company's data and check the prediction accuracy. For those companies where availability of twitter data is a challenge, we would be using yahoo finance news data for similar analysis. We can also incorporate similar strategies for algorithmic trading.

VIII. Conclusions

Our results are in some conjunction with [1], but there are some major differences as well. Firstly our results show a better sentiment analysis of varied languages tweets. Part of Speech tagging makes it a specially fast solution for lexicon-based sentiment classifiers. The classifier engine is implemented in such a way that the presence of absence of n-grams in the terms lists is checked through look-ups on hashsets (is this n-gram contained in a set?), not loops through these sets. Since look-ups in hashsets are typically of O(1) complexity [9].

It’s worth mentioning that our analysis doesn’t take into account many factors. Firstly, our dataset doesn’t really map the real public sentiment, it only considers the twitter using people. It’s possible to obtain a higher correlation if the actual mood is studied. It may be hypothesized that people’s mood indeed affect their investment decisions, hence the correlation.

References

[1] J. Bollen and H. Mao., “Twitter mood as a stock market predictor,” *IEEE Computer, vol. , no. 44(10):91–94*

[2] Anshul Mittal and Arpit Goel Stanford University., “Stock Prediction Using Twitter Sentiment Analysis,” *https://pdfs.semanticscholar.org/4ecc/55e1c3ff1cee41f21e5b0a3b22c58d04c9d6.pdf*

[3] Anurag Nagar, Michael Hahsler, “Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live News Streams,” *IPCSIT vol., no. XX (2012) IACSIT Press, Singapore*

[4] W.B. Yu, B.R. Lea, and B. Guruswamy, “A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting,” *International Journal of Electronic Business Management, vol., no. 2011,5(3): 211-224,*

[5] J. Bean, “R by example: Mining Twitter for consumer attitudes towards airlines,” *In Boston Predictive Analytics Meetup Presentation, Feb 2011*

[6] Yauheniya Shynkevich, “T.M. McGinnity, Sonya Coleman, Ammar Belatreche, Predicting Stock Price Movements Based on Different Categories of News Articles,” *2015 IEEE Symposium Series on Computational Intelligence*

[7] A. Lapedes and R. Farber, "Nonlinear Signal Processing Using Neural Networks: Prediction and System Modeling", *Technical Report, LA-UR87-2662, Los Alamos National Laboratory, Los Alamos, New Mexico, 1987.*

[8] T. Rao and S. Srivastava, "TweetSmart: Hedging in markets through Twitter," *2012 Third International Conference on Emerging Applications of Information Technology*, Kolkata, 2012, pp. 193-196. doi: 10.1109/EAIT.2012.6407894

[9] Clement Levallois, Umigon. Retrieved from *http://www.clementlevallois.net/download/umigon.pdf*