Sentiment Analysis of News Headlines
For Stock Trend Prediction

Oshi Gupta

Student of Bachelor of Technology
Department of Computer Science and Technology
Symbiosis University of Applied Sciences, Indore, India

Abstract: **Stock market is the bone of fast emerging economies. Stock market data analysis needs the help of artificial intelligence and data mining techniques. The volatility of stock prices depends on gains or losses of certain companies. News articles are one of the most important factors which influence the stock market. This project is about taking non quantifiable data such as financial news articles about a company and predicting its future stock trend with news sentiment classification. Assuming that news articles have impact on stock market, this is an attempt to study relationship between news and stock trend.**

Index Terms**: Text Mining, Sentiment analysis, Naive Bayes, Random Forest, Stock trends.**

1. INTRODUCTION

News has always been an important source of information to build perception of market investments. As the volumes of news and news sources are increasing rapidly, it's becoming impossible for an investor or even a group of investors to find out relevant news from the big chunk of news available. But, it's important to make a rational choice of investing timely in order to make maximum profit out of the investment plan. And having this limitation, computation comes into the place which automatically extracts news from all possible news sources, filter and aggregate relevant ones, analyse them in order to give them real time sentiments.

Stock market is the bone of fast emerging economies such as India. Major of capital infusion for companies across the country was made possible only thru shares sold to people. So our country growth is tightly bounded with the performance of our stock market. Considering the fact of lack of knowledge and awareness across the people stock market prediction techniques plays a very crucial role in bringing more people into market as well as to retain the existing investors. There is a strong yet complicated relation between the market and the information available in the form of news. The arrival of news at every moment changes the perception or sentiment towards a particular company or their adopted business strategies. These days due to the bliss of internet, the traders, and investors have constant access to the updated news, and the news constantly mould their sentiments and influences their decision to invest in a particular company. News has always been an important source of information to build perception of market investments. As the volumes of news and news sources are increasing rapidly, it's becoming impossible for an investor or even a group of investors to find out relevant news from the big chunk of news available. But, it's important to make a rational choice of investing timely in order to make maximum profit out of the investment plan. And having this limitation, computation comes into the place which automatically extracts news from all possible news sources, filter and aggregate relevant ones, analyze them in order to give them real time sentiments to know whether stock will go high or down. To discover future trend of a stock by considering news articles about a company as prime information and tries to classify news as good (positive) and bad (negative). If the news sentiment is positive, there are more chances that the stock price will go up and if the news sentiment is negative, then stock price may go down.

This research is to check the impact of news articles on stock prices. I am using supervised machine learning as classification and other text mining techniques to check news polarity. And also be able to classify unknown news, which is not used to build a classifier. Different classification algorithms are implemented to check and improve classification accuracy. I have taken past data of more than five years to do this research.

2. LITERATURE SURVEY

Stock price trend prediction is an active research area, as more accurate predictions are directly related to more returns in stocks. Therefore, in recent years, significant efforts have been put into developing models that can predict for future trend of a specific stock or overall market. Most of the existing techniques make use of the technical indicators. Some of the researchers showed that there is a strong relationship between news article about a company and its stock prices fluctuations.

Following is discussion on previous research on sentiment analysis of text data and different classification techniques.

Nagar and Hahsler in their research [1] 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 NewsSentiment, utilizing the count of positive and negative polarity words is proposed as a measure of the sentiment of the overall news corpus. They have used various open source packages and tools to develop the news collection and aggregation engine as well as the sentiment evaluation engine. They also state that the time variation of NewsSentiment shows a very strong correlation with the actual stock price movement.

Yu et al [2] 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 [3] 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.

I have used the sentiment detection algorithm based on this research.

This research paper [4] studies how the results of financial forecasting can be improved when news articles with different levels of relevance to the target stock are used simultaneously.

3. METHODOLOGY

3.1. System Design

Following system design is proposed in this project to classify news articles for generating stock trend signal.

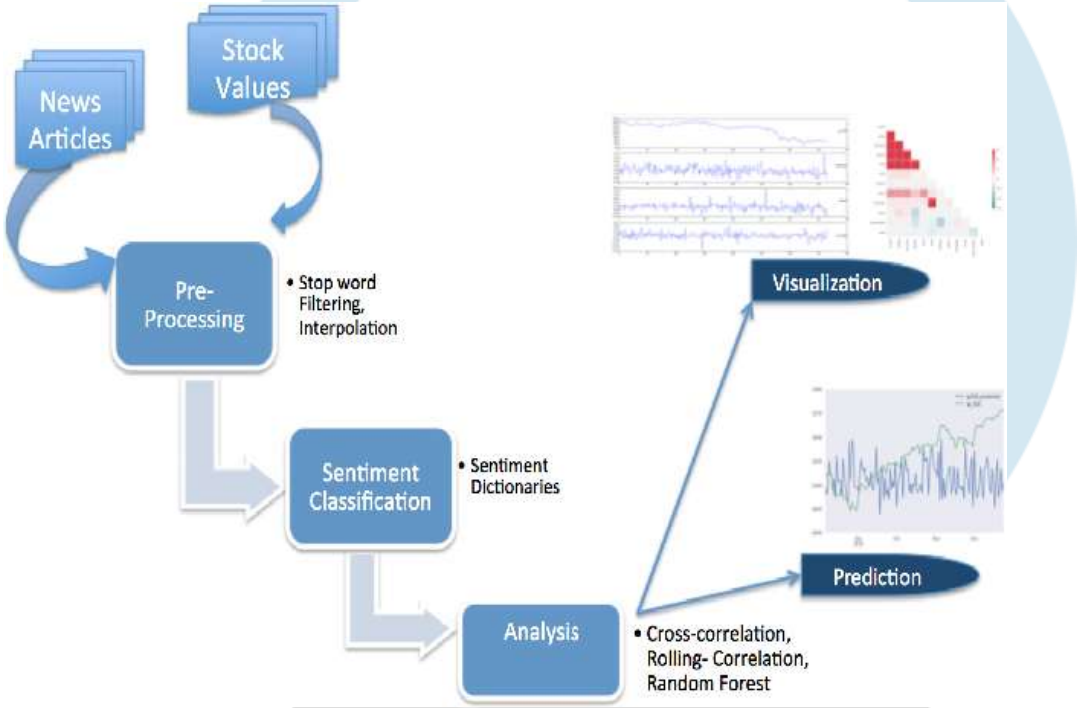

Figure 1: System Design

This design can logically be seen in three parts, part one has news articles dataset with stock values.This dataset is preprocessed to remove punctuation marks, tags. Part two comprises sentiment classification model where after pre-processing, dataset is loaded and different machine learning algorithms applied on it to find which suits it better. After this part three comes where prediction and visualization has to be done of different machine learning algorithms to find which gives better accuracy.

3.1.1. News Collection

I have collected news article data of different countries of more than five years from DJIA (Dow Jones Industrial Average**)** of different countries to analysis stock trend.

3.1.2. Pre Processing

Text data is unstructured data. So, one cannot provide raw test data to classifier as an input. Firstly, we need to tokenize the document into words to operate on word level. Text data contains more noisy words which are not contributing towards classification. So, there was need to drop those words. In addition, text data may contain white spaces, tabs, punctuation characters etc. I also need to clean data by removing all those words. For this purpose, I had used regular expression (regex) to remove punctuation marks and converted the news statements into lower case.

3.1.3. Sentiment Detection Algorithm

For automatic sentiment detection of news articles, I am following Dictionary based approach which uses Bag of Word technique for text mining.

This method is based on the research of J. Bean in his implementation of Twitter sentiment analysis for airline companies.

To build the polarity dictionary, we need two types of words collection; i.e. positive words and negative words. Then we can match the article's words against both these words list and count numbers of words appears in both the dictionaries and calculate the score of that document.

We created the polarity words dictionary using general words with positive and negative polarity. Also addition to this, we used Finance specific words with its polarity using McDonald's research.

For the news article, we are considering the string which contains headline and news body, both.

The algorithm to calculate sentiment score of a document is given below.

Algorithm:

1. Tokenize the document into word vector.

2. Prepare the dictionary which contains words with its polarity (positive or negative)

3. Check against each word weather it matches with one of the word from positive word dictionary or negative words dictionary.

4. Count number of words belongs to positive and negative polarity.

5. Calculate Score of document = count (pos.matches) – count (neg.matches)

6. If the Score is 0 or more, we consider the document is positive or else, negative.

I had done this with Random Forest and Naïve Bayes Classification algorithm.

3.1.4. Document Representation

As in my dataset news articles were splitted into twenty-five columns and to detect stock trend I had to combine these twenty-five column into one column for each record in dataset.

In order to reduce the complexity of text documents and make them easier to work with, the documents has to be transformed from the full text version to a document vector which describes the contents of the document. To represent text documents, we are using TF-IDF scheme. The higher tf-idf value a term gets, the more important it is. A high value is reached when the term frequency in the given document is high and when there are few other documents in the collection containing the given term/feature. This term weighting method tends therefore to filter out common terms by giving them a very low value. Along with this Count Vectorizer is also used to see the difference between this both.

Count Vectorizer is a great tool provided by the scikit-learn library in Python. It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text. This is helpful when we have multiple such texts, and we wish to convert each word in each text into vectors (for using in further text analysis).

3.1.5. Classifier Learning

As most of the research shows that Random Forest and Naïve Bayes classification algorithms performs good in text classification. So, I am considering all two algorithms to classify the text and check each algorithm's accuracy. Also results can be compare on the basis of accuracy, precision, recall and other model evaluation methods. All the classification algorithms are implemented and tested using Weka tool.

3.1.6. System Evaluation

I had divided the data into train and test set. I have evaluated all classifiers performance by checking each one's accuracy, precision, recall, f1-score. The results are as given in the next section.

3.1.7. Testing with new Data

News articles from Dec 2014 to Dec 2016 are used as unknown test set. When comparing results of all classifiers,

Random Forest algorithm with Count Vectorizer and Naïve Bayes algorithm both worked good comparing to Random Forest algorithm with TF-IDF Vectorizer.

4. EVALUATION

The results of algorithms are as follows:

	Accuracy	Precision	Recall	F1-Score
Random Forest with Count Vectorizer	0.851	0.992	0.704	0.823
Random Forest with TF-IDF Vectorizer	0.843	0.899	0.768	0.828
Naïve Bayes	0.851	1	0.698	0.822

Table 1: Evaluation

5. CONCLUSION

Finding future trend for a stock is a crucial task because stock trends depend on number of factors. Assuming that news articles and stock price are related to each other. And, news may have capacity to fluctuate stock trend. So, to study thoroughly this relationship and concluded that stock trend can be predicted using news articles and previous price history.

As news articles capture sentiment about the current market, we automate this sentiment detection and based on the words in the news articles, I can get an overall news polarity.

If the news is positive, then we can state that this news impact is good in the market, so more chances of stock price go high. And if the news is negative, then it may impact the stock price to go down in trend. I have used polarity detection algorithm for initially labelling news and making the train set. For this algorithm, dictionary based approach was used. The dictionaries for positive and negative words are created using general and finance specific sentiment carrying words.

Then pre-processing of text data was also a challenging task. Based on this data, we implemented two classification models and tested it.

Then after comparing their results, Random Forest worked very well with Count Vectorizer with 85% accuracy. Naïve Bayes algorithm performance is also of 85% accuracy. But I will go with Naïve Bayes as its precision was 100% as compared to Random Forest with Count Vectorizer.

Also Random Forest with TF-IDF Vectorizer gives 84% accuracy which is also close to above one.

Given any news article, it would be possible for the model to arrive on a polarity which would further predict the stock trend.

6. FUTURE WORK

I would like to extend this research by adding more company's data and check the prediction accuracy. For those companies where availability of financial news is a challenge, I would be using twitter data for similar analysis and to apply different deep learning techniques to improve their performance.

7. ACKNOWLEDGEMENT

Authors would like to thank our guides, teachers, family and friends who supported in the completion of this research project. Appreciating everyone who helped us knowingly or unknowingly for this project.

REFERENCES

[1] Anurag Nagar, Michael Hahsler, Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live News Streams, IPCSIT vol. XX (2012) IACSIT Press, Singapore

[2] 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. 2011, 5(3): 211-224

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

[4] 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

[5] P. Hofmarcher, S. Theussl, and K. Hornik, Do Media Sentiments Reflect Economic Indices? Chinese Business Review. 2011, 10(7): 487-492

[6] R. Goonatilake and S. Herath, The volatility of the stock market and news, International Research Journal of Finance and Economics, 2007, 11: 53-65.

[7] Spandan Ghose Chowdhury, Soham Routh , Satyajit Chakrabarti, News Analytics and Sentiment Analysis to Predict Stock Price Trends, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3595-3604

[8] Robert P. Schumaker, Yulei Zhang, Chun-Neng Huang, Sentiment Analysis of Financial News Articles

Authors

Oshi Gupta

Student of Bachelor of Technology
Department of Computer Science and Technology
Symbiosis University of Applied Sciences, Indore, India