International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

STOCK MARKET PREDICTION BASED ON SENTIMENT ANALYSIS USING MACHINE LEARNING

Geethu M S[1], Nisha A[2]

[1]*Student Master of Computer Application, Dept. of MCA, Haji C.H.M.M College for Advanced Studies*
[2]*Assistant Professor, Dept. of MCA, Haji C.H.M.M College for Advanced Studies*

Abstract - *Machine learning and artificial intelligence techniques are being used to solve many real world problems. These techniques are highly effective, minimal effort and saving huge amount of time. Now people are invested in stock market or share market for yielding huge amount of money. Stock market is association of buyers and sellers. But in stock market, any time the stock value will grow or down according to the economic trends. So these changes could affect your share value and sometimes that decreases your profit. So stock market prediction is very necessary for avoiding this loss. We will propose a system that predicts the stock market value based on social sentiments using machine learning. We will collect the tweets from twitter API to perform sentiment analysis and at same time collect data from yahoo API. Then find the correlation of historical data and extracted twitter data. This relational value used to determine the predicted outcome. This prediction system could greatly help stock investors in taking desired decision which could affect the profit of stock.*

Key Words: machine learning, sentiment analysis, twitter API, yahoo API, stock market prediction.

1. INTRODUCTION

Stock Market Prediction System (SMPS) is a practical system that forecasts the stock price movement of various companies. Such a prediction could greatly help stock investor in taking desired decision which would directly contribute to his profits. Nowadays, social media has become a mirror that reflects people's thoughts and opinions to any particular event or news. Any positive or negative sentiment of public related to a particular company can have effect on its stock prices. Our system predicts the stock market prices of various companies by performing sentiment analysis of the social media data such as tweets related to the respective companies.

We will collect the tweets from twitter API and perform sentiment analysis of it. Corresponding to that time period, we shall analyze the stock values from past data and use a suitable machine learning algorithm to justify a valid correlation between the tweet sentiment and the stock values. Finally, with training data, we will train our model and develop capability to produce stock predictions for future. Since the public reaction to any event is available almost instantaneously on any social media, their mood can be captured quickly and an estimate of the volatility in stock prices can be determined.

2. RELEVANCE OF THE PROJECT

Stock market prediction is basically defined as trying to determine the stock value and offer a robust idea for the people to know and predict the market and the stock prices. It is generally presented using the financial ratio using the dataset. Thus, relying on a single dataset may not be sufficient for the prediction and can give inaccurate result. Hence, we are contemplating towards the study of machine learning with various datasets integration to predict the market and the stock trends. The movement in the stock market is usually determined by the sentiments of thousands of investors. Stock market prediction, calls for an ability to predict the effect of recent events on the investors. These events can be political events like a statement by a political leader, a piece of news on scam etc. It can also be an international event like sharp movements in currencies and commodity etc. All these events affect the corporate earnings, which in turn affects the sentiment of investors. It is beyond the scope of almost all investors to correctly and consistently predict these parameters. All these factors make stock price prediction very difficult. Once the right data is collected, it then can be used to train a machine and to generate a predictive result.

3. PRESENT SYSTEM

One of the significant financial subjects that have engrossed the researcher's attention for many years is forecasting the stock returns. Investors in the stock market have been attempting to discover an answer to estimate the stock trend in order to decide the timing to buy or sell or hold a share. Forecasting the stock trend has been done both on qualitative analysis and quantitative analysis. There are many statistical models available for forecasting stock trend and choosing an appropriate model for a particular forecasting application depends on the format of the data.

4. DRAWBACKS OF PRESENT SYSTEM

- The previous results indicate that the stock price is unpredictable when the Traditional classifier is used.
- The existing system does not perform well when there is a change in the operating environment
- It doesn't focus on external events in the environment, like news events or social media.
- The existing system needs some form of input interpretation, thus need of scaling.
- It doesn't exploit data pre-processing techniques to remove inconsistency and incompleteness of the data.

5. PROPOSED METHODOLOGY

In this proposed system, we focus on predicting the stock market prices of various companies by performing sentiment analysis of the social media data such as tweets related to the respective companies. The proposed model can be summarized in the following modules:

5.1 Data Collection

First step is tweet collection, we have used search API. The search API is REST API which allows users to request specific query of recent tweets. The search API allows queries filtering based on time, region, language etc. The request of JSON object contains the tweet and their metadata. It includes information including username, time, location, re-tweets. We have focused on time and tweet text for further analysis. An API requires the user have an API key authentication. The text of each tweet contains too much word that do not consider to its sentiment. Tweets include URLs, tags to others and many other symbols that do not have any sentiment value. To accurately obtain tweet's sentiment we need to filter noise from tweet.

First step is to split the text by space, forming a list of individual words which is called a list of words. We will use each word in tweet as feature to train our classifier.

Next, we remove stop words from list of total words using python's Natural Language Toolkit (NLT), which do not have any sentiment value.

Now, tweet contains extra symbols like "@", "#" and URLs. The word next to "@" symbol is always a username which does not have any sentiment value to text tweet.

Words following "#" are kept as they contain information about tweet. URLs are filtered out as they do not have any sentiment meaning. To complete all these processes, we use regular expressions that match these symbols.

5.2 Feature Extraction

After gathering tweet corpus, we have built classifier and train for tweet sentiment analysis. We examine mainly two classifiers: Naïve Bayes and Support Vector Machine. To build feature set, we process each tweet and extract meaningful feature and create feature. The feature set larger and larger as dataset increases. After certain point, it becomes difficult to handle larger dataset. In this case it is not necessary to use every unigram as feature vector to train Naïve Bayes classifier and Support Vector machine. To avoid critical situation, we decided to use 'n' significant feature for training. We have fined the n best features from larger set. It scores each word of training data and distinct n best feature. After calculating feature score, we rank the feature with score and choose top n feature for training and classification. We find tweet sentiment value for training datasets labels.

5.3 Training Data

The generated data is used as training dataset for train the model. On test dataset, we receive the tweet sentiment labels as an output value. We will predict the stock market value using the dataset. We calculate total available stock tweets of each company and generate another dataset which contains positive, negative, neutral and total tweets of each day as a feature matrix. On other side we have taken stock market historical data using Python's yahoo-finance library for each day and have calculated market up and down direction and took it as label for dataset.

5.4 Prediction

After training our classifier, we will check the correlation between tweet sentiment and stock market prices on each day scale. We have collected stock data and tweet data for same timeline. We focus on specific company stocks gathered daily data. After justifying a valid correlation, we are able to predict the stock values.

6. SYSTEM MODEL

The proposed system model consists with following:

- **User :** User is a person who utilize services
- **Input System:** System that is used again the input from various sources. In this the URL and tweets are extracted from newspapers and social media. Real time input streams are gathered in this step.
- **Yahoo API:** The yahoo API is used to collect stock related news from the yahoo finance website. Her URL of the news feeds are gathered and fetch them as a batch file. The extracted data are passed as a plain text from the yahoo feed to our proposed system. Yahoo provides RSS feeds for each news content, they are light weight and easy to process
- **Twitter API:** The twitter API provides tweets with respect to the input keywords. Twitters API's are worked with the twitter are created by the system. in the app the connection parameters are the Authentication key, Access token and the Access Secret. They are highly secured and encrypted. The tweets are generated as JSON files, which contains the tweets content, date and time, user and the location. These files are directly stored in the processing batch
- **Cloud/ Database:** Cloud computing is a type of computing that relies on sharing computing resources rather than having local servers or personal devices to handle applications. In cloud computing, the word cloud is used as a metaphor for the Internet .cloud computing means " a type of Internet based computing " .Virtualization techniques is the main concept used by cloud computing, which is used to maximize the power of cloud computing.
- **SVM:** In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyses data used for classification and regression analysis.

7. ADVANTAGES OF PROPOSED SYSTEM

- It guides people who possess limited know-how of investments and finance into making well informed decisions regarding stock market investments.
- Stock market trends for a given time frame can be analysed easily even by the uninformed.
- A little effort is enough for project to assist the inexperienced investors and prevent from suffering heavy capital loss.
- It cross over the need for hiring investment experts who command exorbitant wages to guide our financial decisions by providing a simple solution which can be accessed by anyone having a computer or a laptop and an internet connection.

8. CONCLUSION

Using Machine Learning technique and sentiment analysis for prediction purposes is inexpensive compared to other models. Support Vector Machine provide to be the most efficient and feasible model in predicting the stock price value. Cloud services will enable us to collect large amount of data and also store it in real time when we will get the data directly from the REST API. Collection of tweets and classification of tweets as positive, negative and neutral gives a good overview of public mood. The proposed system will produce the efficient and accurate results that help stock investor in taking informed decisions.

REFERENCES

[1] Max Sorto, Cheryl Aasheim and Hayden Wimmer,Feeling The Stock Market: A Study in the Prediction of Financial Markets Based on News Sentiment.,2017.

[2] Alostad H,Davulcu H(2015) Directonal prediction of stock prices using breaking news on Twitter,In:IEEE/WIC/ACM international conferences on WI-IAT 1,pp 523-530.

[3] Afzal H Mehmood K (2016) Spam filtering of bi-lingual tweets using machine learning.In:IEEE 18th international conference on ICACT,pp 710-714.

[4] Attigeri GV, MM MP,Pai RM,Nayak A(2015) Stock market prediction:a big data approach.In:IEEE region 10 conference on TENCON,pp 1-5.