PREDICTING STOCK PRICE USING SENTIMENTAL ANALYSIS THROUGH TWITTER DATA

[1]Ladi Eswararao, [2]Pappala Yashaswini, [3]Neelapu Rohit Reddy
Department of Computer Science and Engineering, Raghu Engineering College, Visakhapatnam
[1]19981a0584@raghuenggcollege.in, [2]19981a05b6@raghuenggcollege.in ,
[3]19981a05a4@raghuenggcollege.in
N.SRINIVAS RAO*, Assistant Professor, Department of CSE, Raghu Engineering College, Visakhapatnam, A.P., India.

Abstract

Artificial intelligence and Machine learning methods in combination with data mining are used in multiple scenarios to solve many problems. These Machine learning methods and techniques have already proved to be effective, highly accurate and it saves a lot of time. In recent days, people have started investing in stocks and shares as it is a profitable option in order to increase one's income. If there are proper planning and good guidance there are chances of doubling the annual revenue from the returns we get from the stock market. But even in the present day's people think stock investments still remain a risky theory. Investment experts have very high income along with the ignorance of the general public with respect to the financial problems, some issues like these behave as barriers for many people to invest in stocks. The anxiety of losing the invested money also behaves as a barrier to the people. These facts are the motivation factors for applying the capacity of machine learning to do the prediction on the movements of stocks. The sentimental analysis is used on the tweets which are obtained by using the Twitter API. Such forecasts are of great use for stock investors so that they can take calculative decisions and invest in stocks that are profitable.

Keywords: Artificial intelligence, Twitter API, sentiment analysis, machine learning.

1. Introduction

Predicting price of stock market is a tedious task. There are two methods, they are technical analyses and fundamental analyses. Stock market price prediction is a monotonous task, this is solely because stock time series behaves as a close to random walk. The companies have to hiring investment experts who will take unreasonably high income in order to council about financial decisions. To build an efficient, inexpensive and feasible model to forecast the stock market price, this is done by applying sentiment analysis on twitter data. These facts are the motivation factors for applying the capacity of machine learning to do the prediction on the movements of stocks. We need to minimize the difference between predicted values from the model and the actual values. In all most all thee-commerce websites and social networking applications comments, text reviews, feedbacks can be provided by the 978-1-7281-6828-9/20/$31.00 ©2020 IEEE users. This user generated texts are a rich source of user's sentiment opinions which are based on multiple items and products, these opinions are used for building recommender systems.

1.1 Motivation

The main goal of this paper is to educate people who do not have knowledge on finance and investments regarding the stocks. It detours the needs of the investment on experts who have very high income along with the ignorance of general public with respect to the financial problems, some issues like these behave as barriers for many people to invest in stocks. Trends in Stock market can be analyzed by common man when a particular amount of time is given. Machine learning is considered as an option to provide cheap alternatives to many stock market investment guidance agencies which are trending nowadays. Therefore, small efforts have been being put to assist people who are inexperienced in investments.

1.2 Objectives

•To build an efficient, inexpensive and feasible model to forecast the stocks.
•To minimize the difference between predicted values from the model and the actual values.
•To feed the Machine learning model and to plot the corresponding results.

2. Literature Survey

There are two approaches for the prediction of stock market, technical and fundamental approach. In technical approach uses the historical data such as: volatility, volume of trading, past training etc., whereas the experimental approach is used for external information like stock prices, interest rates.

The methodology used is SVM and sentiment analyzer, in this scenario Support Vector Machine is used in various projects it have been proved to be the most easy and efficient model for the prediction of variations in stock price, based on the preference of sentiments of the tweets. Since it is based on user review the fake reviews may influence the model output. A machine learning approach which is used to solve a plethora of problems in the recent years some of the problems are sudden financial crashes in the stock .In order to get good results regarding prediction of the stock price linear classifiers are used or the same process can be done by using the S&P500 index.

The methodology used is LSTM, the results obtained by using Long short-term memory networks has been proved to be a path which is assured for the prediction of stock. This technique verifies the effectiveness of a types of Recurrent on the usage of AAPL ticker from NASDAQ exchange so that prediction of stocks prices that is useful for LSTM are easy. Adam optimizer is helpful for testing the three popular output activation layers as back-propagation algorithm. Deviation of root mean square is used for the comparison of the performance. The limitations when we use LSTM as our model is poor data quality, serious data loss and faked data. The methodology used is Artificial Neural Networks (ANN) in MATLAB. The result of using Artificial Neural Networks is that the predicted price was too accurate with respect to the actual price. The limitation is that, since it will be built in MATLAB, we can’t use it in small devices like a mobile phone. The methodology used is Convolutional Neural Network (CNN) model. Where the neural network will do the detection of complex features. This accuracy of this model is almost 50%, this accuracy is better when compared to the random probability which has highest accuracy recorded by the model is 53%.

Bollen explains about the public mood there are three types of public moods they are happy, calm, anxiety, firstly we correlate the collective mood of the public which are derived from the feeds from twitter to the Dow Jones Industrial Index values. Dow Jones Industrial Index value is called as Stock market index which helps in measuring the stock performance of the thirty companies which are listed in the stock exchanges in United States. Fuzzy neural networks are used in this model for prediction. The results from this tell that the public moods and Dow Jones Industrial Index values are strongly mapped and correlated to each other . The complicated techniques include the support vector machine which is also called as large margin classifier, this technique was the best before the theory of neural network existed. This uses a kernel trick which is useful for the consideration of large amounts of input data into higher dimensional space.

In this higher dimensional space, it can be linearly separable. For example, we can consider the stock price market of the Korean composition in.Using Pearson correlation coefficient we can predict the increase and decrease in the stock market .The methodology used is Logistic Regression, Decision Tree, SVM and KNN the model is trained with multiple algorithms to increase the accuracy but the drawbacks found by trained this are the visualization, when there are large datasets the results are not accurate, the model is very sensitive if there is any noisy data and results were not proper .The technique used is the sentimental analysis algorithm in which the sentiment of each review and each sentence is calculated. The major drawbacks of this approach are the fewer features were extracted which has decreased the model accuracy .

Sentiment analysis and finding the correlation between them but significant results were not obtained.Another methodology where the method used is NLTK, sentimental-tool from GitHub in which the sentiment polarity score of the news headline is used to correlate the price and sentiment but there is Poor ROC values about 0.75, the ROC is referred to as the rate of the change in the price from one instance of time period to the next time period.The artificial neural network is used where the predicted values were almost near to the actual values trained with the old data which is not able to predict today's price properly.

Another method which is used is recurrent neural network and convolution neural network implemented a novel Wavelet and also CNN algorithm which overcomes the other neural network approaches. The major drawbacks are the accuracy of the model is very low in terms of accuracy and overfitting.

	S&P501	Forex
MLP (101-51)	0.69	0.48
MLP (501-251)	0.55	0.47
CNN (2 layers)	0.59	0.44
RNN (2 layers)	0.59	0.38
Wavelet CNN	**0.54**	**0.44**

Figure1: Log-loss results after training the different architectures.

	S&P501	Forex
MLP (101-51)	0.58	0.82
MLP (501-251)	0.7	0.84
CNN (2 layers)	0.60	0.83
RNN (2 layers)	0.64	0.77
Wavelet CNN	**0.68**	**0.85**

Figure 2: Accuracy after training the different architectures.

Therefore, the above tables give the results of the using convolution neural networks.

Supervised learning algorithm can be used, by using this there was 80% model test accuracy and the issues was regarding Visualization, and results were not proper. The assessment of cross validation process is conducted to transform into time series, there are three main criteria to analyze the performance: percentage of opportunities seized by the classifier, percentage of successful operations advised by the classifier and average return per operation. Word2vec and N-gram can be used for prediction of stock price, by using this method the rise and fall in stock market can be predicted and the outcome of this can be compared and co related with the opinion of the pubic which can be collected using twitter . However, the limitations of using Word2vec is that it cannot be optimized dynamically if there are any specific tasks need to be done. Word2vec is a static model and majorly N gram is only useful when there is huge amount of data. Sometimes even if the model has a huge dataset it cannot represent every unseen instance in the feature space.

3.Proposed Methodology

In today's world Stock market has become one of the big investments. This is because the stock market has the capacity of advancing and optimizing capital allocation, financing and also helps in increasing the value of properties. For a day trader, investment experts or financial advisers are majorly based on the prediction of stock price. They can make their next move based on the prediction. Twitter is the platform for us to mine the database for prediction. This model also combines with the sentiment analysis used to predict the value of the stock. The core of the prediction method is the Machine learning model that learns from the past data to make predictions. Stock price prediction is solely a repetitive and tedious task because stock time series behaves as a close to random walk. The companies have to hiring investment experts who will take unreasonably high income in order to council about financial decisions. Fear of losing the investments act as a barrier to many people. These facts are the motivation factors for applying the capacity of machine learning to do the prediction on the movements of stocks. The sentimental analysis is used on the tweets which are collected from twitter, we obtain tweets with the help of Twitter API and model a system which can predict stock price movement for various companies. In order to analyze and predict based on the public mood using sentimental analysis, support vector machine is an effective model. The design and framework for stock market prediction can be done as shown in figure 3.

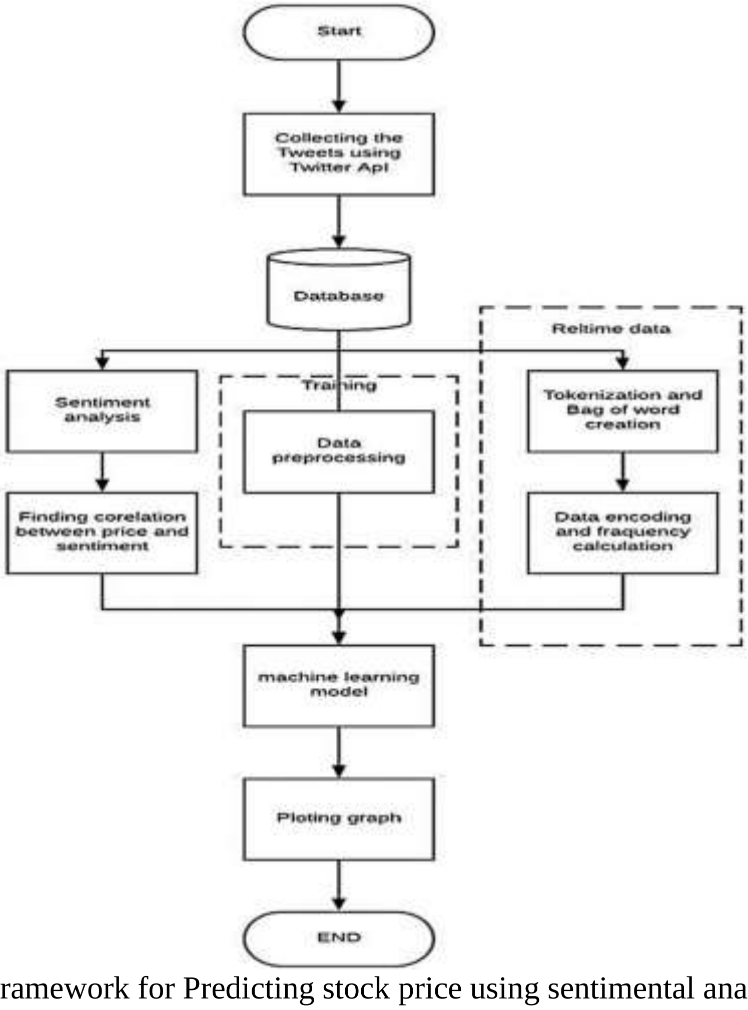

Start
Collecting the Tweets using Twitter Api
Database
Sentiment analysis
Finding corelation between price and sentiment
Training
Data preprocessing
Reltime data
Tokenization and Bag of word creation
Data encoding and fraquency calculation
machine learning model
Ploting graph
END

Figure 3: Proposed framework for Predicting stock price using sentimental analysis through twitter data.

3.1 Data Collection

To collect the tweets from twitter, a robust API is there in twitter. To gathering of tweets there can be two possibilities: search API and streaming API. Search API is the REST API using this API specific queries can be requested from recent tweets. The fine tuning of the queries can be filtered based on region, time, language etc using search API. JSON object request contains the tweet and metadata. For analysis purpose we focus on tweet text and time. The user must have the API key authentication, which is required by the API. The python libraries are called Tweepy. These python libraries can be accessed after the authentication is done. The authentication is done by using this authentication key. The text data in a tweet has a lot of extra data which will not be considered for the analysis of sentiments. For this purpose, the tokenization by using bag of words creation, data encoding and frequency calculation.

3.2 Training Module

For training, we will be using the time-series data for stock price prediction. The collected data will be tokenized, and a bag of words is formed. Now we can calculate the frequency and sentiment of each post which can be used as the features for our model. Once the feature is extracted data normalization is done which completes the data preprocessing part. The preprocessed data will be fed to training and the model will be tuned to give better accuracy for the existing data. In the next step the trained model will be saved. The generated data which is used for the training data set which is used for training the model for the sentimental analysis. When the model is inspected based on the test dataset is used for the prediction of stock market. The total number of available stock tweets are calculated with respect to a particular company. Another dataset is generated which contains negative, positive and neutral tweets this dataset also contains total number of tweets per day based on the feature matrix. After his data encoding is the process which is converting alphabets, symbols to a particular format for secures transmission and frequency calculation is done.

3.3 Prediction Module

After he model is trained, the correlation is found between the price and sentiment this value is used for the future predictions. A machine learning model is used for the prediction purpose. The machine learning model is trained based on these values and the prediction will be done. We will use Random forest algorithm for the prediction purpose. In Random Forest algorithm there are two stages, in the first stage random forest creation is done. In the second stage prediction is done based on the random forest classifier built in the first stage.

4. Algorithms:

4.1 Support Vector Machine:

Support Vector Machines is a supervised machine learning algorithm, adopted conventionally for classification as well as regression problems. SVMs for classification, work by figuring out the right hyperplane among the classes. After being trained by a labeled data set, SVM outputs an optimal hyperplane that categorizes new examples. Classification by SVMs for different data sets is governed by tuning parameters namely kernel, regularization, gamma and margin. When data is 2 dimensional Support vector classifier is a line, if it is 3D SVC forms a plane instead of a line. When data is more than 4D then classifier is a hyperplane. For highly distributed data Maximal margin and support vector classifier fail and hence SVMs are used. For linearly separable patterns optimal hyperplane is formed and for non-linearly separable patterns transformation of original data into a new space is performed determined by kernel function. To classify tweets into different emotion classes a linear kernel has been utilized. Linear kernel is preferable for text classification problems because text has lot of features, linear kernel is faster and less parameters are optimized. When SVM is trained with a linear kernel only C regularization parameter need to be optimized whereas for other kernels you need to optimize gamma parameter also.

Linearly Separable Case

If the training data are linearly separable, then there exists a pair (w,b) such that WT Xi + b $\geq$ 1, for all Xi $\in$ P (1)WT Xi + b $\leq$ -1, for all Xi $\in$ N

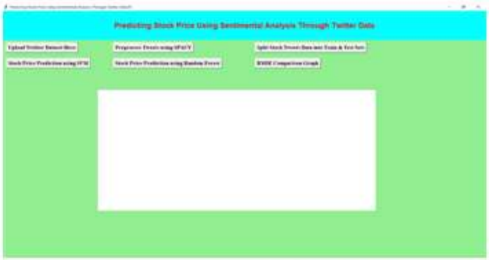
The decision function is of the form f(x) = sign(wtx + b) (2)

4.1 Random Forest

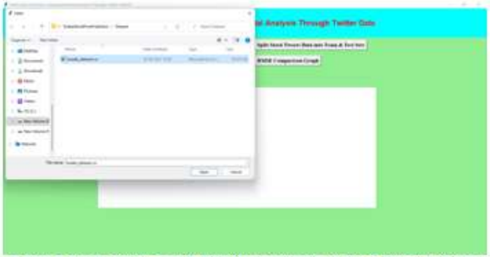
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

5.SAMPLE RESULTS


Predicting Stock Price Using Sentimental Analysis Through Twitter Data
Upload Twitter Dataset Here
Preprocess Tweets using SPACY
Split Stock Tweets Data into Train & Test Sets
Stock Price Prediction using SVM
Stock Price Prediction using Random Forest
RMSE Comparison Graph

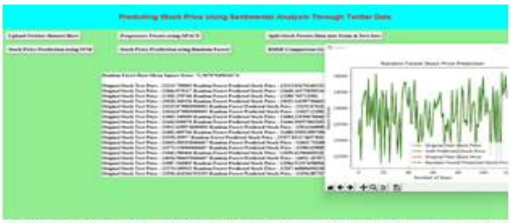
In above screen click on 'Upload Twitter Stock Dataset' button to upload dataset and get below screen



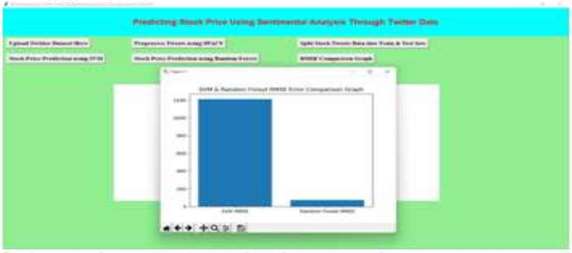
In above screen selecting and uploading tweets stock price dataset and then click on 'Open' button to load dataset and get below output

In above screen tweets and stock prices loaded and now click on 'Preprocess Tweets using SPACY' button to read all tweets and then find sentiment probability of each tweet in terms of positive, negative and neutral and then will get below output

In above screen for tweet we got customer experience in terms of sentiments of stock performance and now click on 'Split Stock Tweets Data into Train & Test'



In above screen with Random Forest we got RSME as just 21 and in graph we can see both lines are fully overlapping so TEST prices and Random Forest predicted prices are accurate. So Random Forest is best in performance
Now close above graph and then click on 'RMSE Comparison Graph' button to get below graph



In above graph x-axis represents algorithm names and y-axis represents error rate and in both algorithms Random Forest got less error so its performance is best. The lower the error the better is the algorithm

6. Challenges

There are few challenges for doing this prediction they are: (1) In order to fetch the real time data from twitter there is an authentication which is required. (2) Out of the huge amount of data sorting out the data which is required. (3) Data from the twitter can be obtained from a certain period of time, historical data is saved by someone.

7. Conclusion

This project proposed a machine learning model uses Random Forest for stock price prediction using Twitter reviews. These reviews include emotions i.e. polarity and the comment about the product. The PSO is employed iteratively as global optimization algorithm to optimize Random Forest for stock price prediction. Also, plot all the data related to results and the training part. Based on the surveys and comparisons done with all the other machine learning models for the stock price prediction using twitter, in order to analyze and predict based on the public mood using sentimental analysis, Random Forest is the most inexpensive model. In order to get an overview of the public mood the tweets are classified into positive, neutral and negative.

References

T. Mankar, T. Hotchandani, M. Madhwani, A. Chidrawar and C. S. Lifna, "Stock Market Prediction based on Social Sentiments using Machine Learning," 2018 International Conference on Smart City and Emerging Technology (ICSCET), Mumbai, 2018, pp. 1-3, doi 10.1109/ICSCET.2018.8537242.

Syed, Shahan & Mubeen, Muhammad & Hussain, Adnan & Lal, Irfan. (2018). Prediction of stock performance by using logistic regression model: evidence from Pakistan Stock Exchange (PSX). Asian Journal of Empirical Research. 8. 10.18488/journal.1007/2018.8.7/1007.7.247.258.

M. A. Asraf Roslan and M. H. Fazalul Rahiman, "Stock Prediction Using Sentiment Analysis in Twitter for Day Trader," 2018 9th IEEE Control and System Graduate Research Colloquium (ICSGRC), Shah Alam, Malaysia, 2018, pp. 177-182, doi: 10.1109/ICSGRC.2018.8657614.

S. Kumar and D. Ningombam, "Short-Term Forecasting of Stock Prices Using Long Short Term Memory," 2018 International Conference on Information Technology (ICIT), Bhubaneswar, India, 2018, pp. 182-186, doi: 10.1109/ICIT.2018.00046.

H. Yun, G. Sim and J. Seok, "Stock Prices Prediction using the Title of Newspaper Articles with Korean Natural Language Processing," 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Okinawa, Japan, 2019, pp. 019-021, doi: 10.1109/ICAIIC.2019.8668996.

Johan Bollen, Huina Mao, Xiaojun Zeng, Twitter mood predicts the stock market, Journal of Computational Science, Volume 2, Issue 1,2011,