ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 8, Issue 9, March 2019

Optimization of stock market values Prediction by using Machine Learning

Shridhar Kumar Thakur, Garima Pandey
Department of CSE Galgotias University, India

Abstract*— In recent past, data mining, artificial intelligence, and machine learning have gained enormous attention in many fields. Stock market is a field where the stock prices have variations. Machine learning (ML) is considered as a significant technique. It will be a high economic advantage if we could give exact predictions of the stock prices. Machine Learning techniques are here used to evaluate past data of the world affairs of a Allotted time period. This work presents some of the possible techniques to predict the stock values with some accuracy. The model we built will be able to buy or sell the stock values on a profitable condition. We used Natural Language processing model to make Smart "decisions" based on the articles and current affairs related to the stock market. With the basic rules of probability and NLP our aim is to make possible accuracy of the stock prediction.

***Index terms*-Machine Learning Algorithm, Feature Scaling, Feature Extraction, Neural Networks, Logistics.**

I. INTRODUCTION

Natural language processing is a technique which is used by computers to understand and make change to the natural languages. Natural language means all the language derived by human. NLP (natural language processing) is used so that machine could analyze the input and derive the meaning. The human computer interaction allows to come up with many different application to bring machine and man as one. we can take Google translator as an example, there NLP works for speech recognition.

In our project we will use some of established NLP techniques to go through the past data of the stock market and the world affairs of the corresponding time period, to make predictions in stock trends.

As to proceed further with this objective, we needed to understand what Sentimental Analysis is. Sentimental Analysis is used to analyse the natural language and deduce if the entered message by user is negative, neutral or positive. In case of our model, Sentimental analysis means, the deduction of the news headlines because they can effect our stock by increasing or reducing it. This will result to the ending of 'emotional' status of the data which is given by sentimental analysis to its user. For collecting and wrangling of data we have used the Combined News DJIA.csv data set. The span was set from 2008-2016 by the DJIA.csv data. To include additional data we had have extended the Data set. We collected some additional data from the

Guardian's Restful News API for the time span of 2000- 2008 period. All the 25 most popular headlines of each day were taken into account of the year 2000-2008 period. For accuracy, From the Finance's website of Yahoo, we took (DJI) for the time span of 2000- 2008 to compare the influence of the available data. The news data containing past news headlines were taken from the "redit world" news channel. The news source is two difference sources. For a single day, the top 25 headlines are taken into account. The Industrial Average Stock Data for Dow Jones (DJIA) is used for marking the time span of the data. The inventory details of Yahoo Finance have been collected.

Note: The explanatory data is derived from the headline of each data that causes the stock price to either rise (labeled 1) or to fall (labeled 0). Form everyday news we took the top 25 headlines and arranged as one row of the extracted data set. Since our goal is to predict the tendency of the stock of a specific company, the data that suggest if the stock's price of the next day to decline or stay the same are labeled "0", while the data that suggests that the price of the stock to rise on the next day are labeled "1". We merged the data from two different pulled data set after comparing them to get the more accurate prediction.

It was not easier to proceed much further if we have not manipulated the given raw data to suit our analysis. We have converted the raw data into vectors. Because, vectors are much easier to work on. For the conversion of the raw data to vector, we used the method "Word2Vec". Word2Vec is used to create word embeddings. In NLP word embeddings are sets of feature learning and language modeling where mapping of phrases or words from the vocabulary vectors of real numbers are done. All the training and test sets are making up by the vectors. Due to the spaces used in the English language it becomes really easy to tokenize – in other words, to recognize what word is given, we used a simple set of rules for English tokenization. With the help of Python, we manipulated the given raw data.

II. PROBLEM STATEMENT

We had to convert the training data set into numeric representation to apply machine learning techniques. The 'Bag of Words' model were used by us because,. This model is used to derive and remember vocabulary from all the given documents, after that documentation is done by counting the number of times each word appears. The values obtained are the feature vectors which are derived from the model. The thing is we cannot stop at just using a lot of Words model as this generates feature vectors that only give importance to the number of occurrences of words, doesn't matter where the word occurred and with

Manuscript received: 23 February 2019
Manuscript received in revised form: 19 March 2019
Manuscript accepted: 07 April 2019
Manuscript Available online: 10 April 2019

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 8, Issue 9, March 2019

what words they accompany. we use the n- gram model or the skip gram model to get past this. Now, with the help of this model, the words can be stored in the same manner as they occur in the data. It depends on "n" that how much words will be stored in a single order. Let n=4, calls for a bigram model it will stores 4 words in this order. To construct the data set we use the Natural Language Processing (NLP) tool. The data are in form of raw of lines (sentence). To reduce the complexity in the given data, the stop words like "a" and, "the" have been removed from the data set. Further, we have used the N-gram model. By the help of this model we are able to predict the next set of words in an n- worded text or speech. To focus on the sentiments of the given words with the bag of words concept we use Google's Word2Vec deep learning method we used this model because, this method doesn't need labels in creation of meaningful representations. Intriguing characteristics word vector would be produced if there is presence of enough training data. It will help us to analyze the relationship between the words with similar meanings. After this implementation, we obtain manipulated data vectors ready to be trained and tested. The extracted dataset is now splited in the ratio of 4:1. 80% of the extracted data will be the training data and 20% of the extracted data will be the test data.

III. PROPOSED MODEL AND METHOIOLOGIES

We are working on four different types of machine learning. The models are Naive Bayes, Random Forest and Logistic regression. We used a jupyter notebook, an open source web application. Jupyter notebook helps us to build and share live code, calculations, views and explanatory documents. We perform cleaning and transformation of data, statistical modeling and machine learning in this environment.

Classification In this area the set of data gets analyzed and categorized under its particular column on the basis of its similar attributes [3,4,8]. Through the datasets or values under analysis this method concludes the process from observed values. In the case of multiple input, the output also expected in various form. Random forest classifier, SVM classifiers are included in this work.

1) Random Forest Classifier: A classifier used beside a supervised algorithm. It yields results on the basis of developing decision trees. The basic approach of this classifier is to proceed with decision aggregate through random subset decision tress as well as gives a final result depends on votes of the random subset of decision trees.

2) Parameters: To mention generalized accuracy, Random forest classifier is included and mentioned n estimators denotes total number of decision tree and for other hyper parameters adobo- score, max_features consists of number of features for the best-split. The minimum weighted fraction of total weights of all the input samples needed for leaf node was obtained by Min_weight_fraction_leaf. . Samples were considered to be of same weight, at the time providing samples without its exact weight. The accuracy in this model obtained is 0.846.

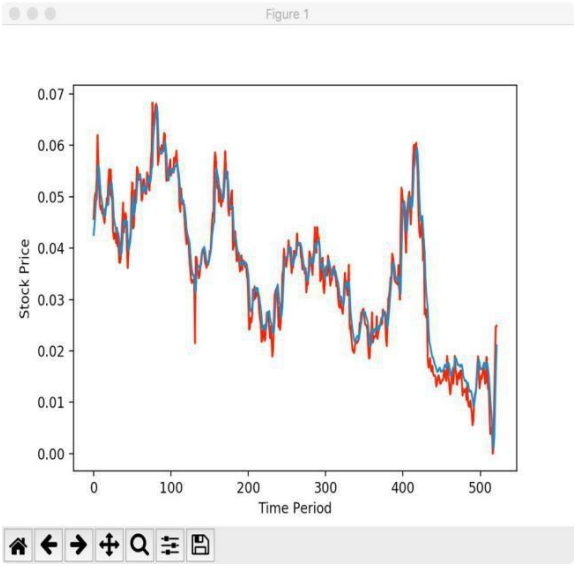
3) SVM classifier: It is a discriminative classifier. Then working of SVM classifier is performed on the basis of supervised learning i.e. a labeled training data. Output is in hyper plane mode to categorize new dataset. They are SVM with the aid of learning algorithm for the purpose of both classification and regression.
The accuracy in this model is 0.854

4) Logistic regression Algorithm: It is used for predicting stock market as it is the easiest and most flexible methods of machine learning algorithm, it leads to exact prediction. The most common use of this algorithm is done in classification tasks.

5) Support Vector Machine Algorithm: Support machine algorithm mainly focuses on identifying an N-dimensional space that differentiates other data points. N denotes the number of features. The output in SVM (Gausian): 0.8492 The output in SVM (Linear): 0.8544.

Between two data points, many hyper plane can be chosen. This algorithm focuses to obtain plane with maximum margin, it refers to the distance between data points of classes under discussion [1, 4, 9]. Benefits like, giving reinforcement for future data points to easy classification is contained. Hyper planes are those decision boundaries which are used to categories data points, on the basis of data points in hyper planes; which are attributed to various classes. Its dimension rests on number of attributes, if it is two and the hyper plane is a line, if it is three the hyper plane is two dimensional[8,10].

IV. EXPERIMENTAL RESULTS



ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 8, Issue 9, March 2019

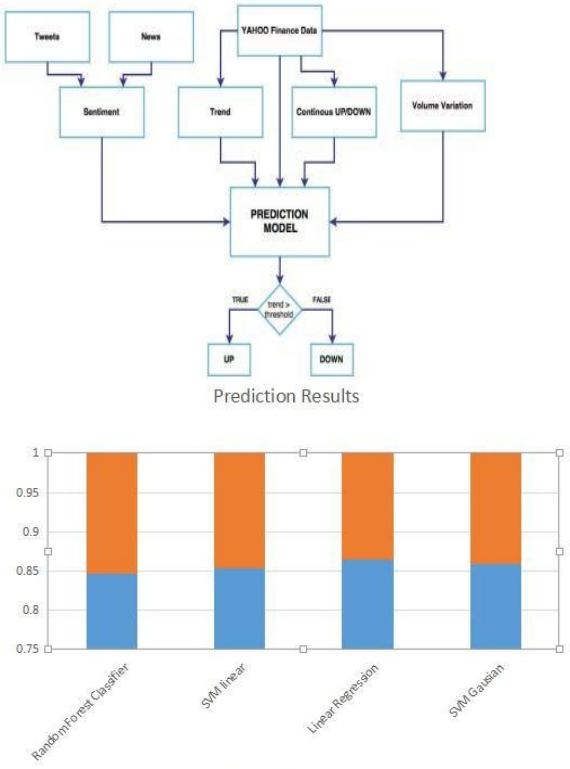


Fig .1. Graph of Prediction Results

```
14]: basicvectorizer3 = CountVectorizer(ngram_range=(2,3))
basictrain3 = basicvectorizer3.fit_transform(headlines)
print(basictrain3.shape)

basicmodel3 = LogisticRegression()
basicmodel3 = basicmodel3.fit(basictrain3, train["Label"])

basictest3 = basicvectorizer3.transform(testheadlines)
predictions3 = basicmodel3.predict(basictest3)

pd.crosstab(test["Label"], predictions3, rownames=["Actual"], colnames=["Predicted"])

print (classification_report(test["Label"], predictions3))
print (accuracy_score(test["Label"], predictions3))
```

```
(3975, 1553543)
              precision    recall  f1-score   support

           0       0.89      0.83      0.86       186
           1       0.85      0.90      0.87       192

avg / total       0.87      0.87      0.86       378

0.865079365079
```

Out[11]:

Predicted	0	1
Actual		
0	143	43
1	17	175

```
In [12]: from sklearn.metrics import classification_report
from sklearn.metrics import f1_score
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix

print (classification_report(test["Label"], predictions))
print (accuracy_score(test["Label"], predictions))
```

```
              precision    recall  f1-score   support

           0       0.89      0.77      0.83       186
           1       0.80      0.91      0.85       192

avg / total       0.85      0.84      0.84       378

0.84126984127
```

Fig .2. Outputs

V. CONCLUSION

Through these Machine Learning algorithms perfect algorithm for the purpose of predicting the stock market value is found on the basis of research conducted with various data from collected data history [1, 7]. After several attempts done on sample data, the specified algorithm which is suitable for exact prediction stock values; it will be useful for both investors and brokers [2, 5]. This research work projects and proceeds with ML method and the value of stock market goods are foreseen finely while comparing with past modules and research works.[2,3] Various things like, Multiple instances, financial ratios, Parameters, etc are included for further research. As per the amount of parameters included in the research process, the prediction of stock market value comes with accuracy. These kind of algorithms are processed to analyse the contents public reviews and patterns as well as the relationships between the customer and the corporate [6]. The best accuracy was obtained in Logistic Regression Algorithm.

REFERENCES

[1] F. Xu and V. Keselj," Collective Sentiment Mining of Micro blogs in 24-hour Stock Price Movement Prediction", IEEE 16th Conference on Business Informatics, Geneva, pp. 60-67,2014.

[2] L. Bing, K. C. C. Chan and C. Ou, "Public Sentiment Analysis in Twitter Data for Prediction of a Company Stock Price Movements", IEEE 11th International Conference on e-Business Engineering, Guangzhou, pp. 232-239, 2014.

[3] D. Rao, F. Deng, Z. Jiang and G. Zhao, Qualitative Stock Market Predicting with Common Knowledge Based Nature Language Processing: A Unified View and Procedure," 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, pp.381-384,2015.

[4] Z. Jiang, P. Chen and X. Pan, "Announcement Based Stock Prediction", International Symposium on Computer, Consumer and Control, pp. 428- 431, 2016.

[5] W. Bouachir, A. Torabi, G. A. Bilodeau and P. Blais, "A bag of words approach for semantic segmentation of monitored scenes", International Symposium on Signal, Image, Video and Communications (ISIVC), Tunis, pp.88-93, 2016.

[6] D. Sehgal and A. K. Agarwal, "Sentiment analysis of big data applications using Twitter Data with the help of HADOOP framework", International Conference System Modeling & Advancement in Research Trends (SMART), Moradabad, pp. 251-255, 2016.

[7] R. Zhao; K. Mao" Fuzzy Bag-of- Words Model for Document Representation", IEEE Transactions on Fuzzy Systems, Volume: 26, Issue: 2, April 2018, pp: 794-804, 2018.

[8] V. U. Thompson, C. Panchev and M. Oakes, "Performance evaluation of similarity measures on similar and dissimilar text retrieval", 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), Lisbon, pp. 577-584, 2015.

[9] C. Sreejith, M. Indu and P. C. R. Raj, "N-gram based algorithm for distinguishing between Hindi and Sanskrit texts", Fourth International Conference on Computing,

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 8, Issue 9, March 2019

Communications and Networking Technologies(ICCCNT), Tiruchengode, pp. 1-4, 2013.

[10] M. Kaya, G. Fidan and I. H. Toroslu, "Sentiment Analysis of Turkish Political News", IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Macau, pp. 174-180, 2012.