International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Stock Prediction: NLP and Deep Learning Approach

Mr. Yogesh Bodkhe and Prof. Rushali Deshmukh

Department of Computer Engineering Rajarshi Shahu College of Engineering Pune,

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

People have a tendency to analyze existing strategies and so planned new strategies for inventory prediction. We have used Sentiment evaluation and Technical evaluation through NLP and Deep mastering approach. In order to exploit benefits of sentiment analysis on enterprise associated inventory, we have proposed a machine that will use the sentiment analysis on twits associated with special sectors (e.g. IT sector, Banking sector, Pharmaceutical sector, Automobile sector, Infrastructure sector.) which might be extracted from twits. These twits are extracted from twitter for calculating polarity. The rating of sentiment analysis is calculated here by using algorithm. According to sector we've taken 20 groups. Top four performer businesses of every sector. Using polarity score we will finalize pinnacle ten groups with great sentiment rating. We will down load the CSV facts of historical share charge of top ten organizations that we've selected. Then downloaded CSV records are used to build a CNN version to predict in addition stock movement of these pinnacle ten companies.

Keywords: *Stock prediction, Natural Language processing, Deep Learning, Price forecasting*

Introduction

Financial analysts investing in stocks usually but they are not aware about the inventory market place conduct. They are going through the problem of trading as they do not properly recognize which shares to shop for or which shares to promote with a purpose to get greater profits. In today's world, all the information pertaining to inventory market is available. Analyzing all this records in my opinion or manually is pretty difficult. As such, automation of the method is required. This is where Data mining techniques help. Understanding that analysis of numerical time series offers close results, wise traders use system learning techniques in predicting the inventory market conduct. This will allow financial analysts to foresee the conduct of the inventory that they may be interested by and consequently act accordingly. The input to our gadget will be historical data. Appropriate data could be carried out to discover the stock fee trends. Hence the prediction value will notify the up or down of the stock price movement for the buying and selling of stocks and traders can act upon it so one can maximize their chances of gaining a profit.

Key Take Always of Deep Learning

- Deep learning is an AI work that mirrors the activities of the human mind in preparing information for use in basic leadership.
- Deep learning can learn from the data that is from both unstructured and unlabeled source.
- Deep learning is machine learning's subset that can be used to help to detect fraud or money laundering.

Literature Survey

Rakhi Batra et al [1]. had performed sentiment analysis on tweets which was related to Apple products, which extracted from Stock Twits (a social networking site) from 2010 to 2017. Along with tweets, along with they had used market index data which was extracted from Yahoo Finance for that same period. The sentiment score of that tweet was calculated by sentiment analysis of tweets through SVM. As a result each tweet was categorized as bullish or bearish. Then they had to used sentiment score and market data to build a SVM model and to predict next day's stock movement. They implemented the idea by collecting sentiment data and stock price market data and built an SVM models for prediction and they had measured the prediction accuracy.

Yaojun Wang et al. [2] had used the social media mining technology to quantitative evaluation market segment and in combination with other factors to predict the stock price trend in short term. Their experiment results showed that by using social media mining combined with other information so the stock prices prediction model can forecast more accurate. Starting from the efficient market hypothesis they had fetched the stock comments information from social

media and then they preprocessed the data to emotion vectors. By calculating the segment value of each stock's emotion vector, they found that the segment value is very sensitivity to the stock price movement.

Ashish Sharma et al. [3] had made survey of well-known efficient regression approach to predict the stock market price from stock market data based. The aim of their research study was to help the stock brokers and investors for investing money in the stock market.

Ze Zhang et al. [4] had taken Elman network to predict the opening price of stock market. Considering that Elman network was limited, so their work adopts self-adapting variant PSO algorithm to optimize the weights and thresholds of network. Afterwards, the optimized data, regarded as initial weight and threshold value was given to Elman network for training, accordingly the prediction model for opening price of stock market based on self-adapting variant PSO-Elman network was formed. At the last their work verified that model by some stock prices and compared with BP network and Elman network. So they had drawn the result that showed precision and stability of that predication model both were superior to the traditional neural network.

Dev Shah et al. [5] had retrieved, extracted and analyzed the effects of news sentiments on the stock market. Their main contributions included the development of a sentiment analysis dictionary for the financial sector, the development of a dictionary-based sentiment analysis model and the evaluation of the model for gauging the effects of news sentiments on stocks for the pharmaceutical market. Using only news sentiments, they achieved a directional accuracy of around 70% in predicting the trends in short-term stock price movement. One major contribution of their work was sentiment analysis dictionary. The sentiment scores obtained from the analysis of the news articles was a powerful indicator of stock movements and that can be used to effectively leverage the prediction of short-term trends.

Du Peng [6] had mainly studied the specific mechanism of investor sentiment affecting stock market volatility.

His work collected the data of web news emotion index, web search volume, social network emotion index, social network heat index and established corresponding analysis index. After correlation analysis and Granger causality tests, it extracted the indicators which had significant correlation with the financial market and brought them into forecasting analysis. The model constructed market volatility index and analyzed the correlation between investor sentiment and stock price changes. Muhammad Firdaus et al. [7] were aiming literature review to explore the use Artificial Neural Network (ANN) techniques in the field of stock market prediction. Their study revealed that the ability of an artificial neural network (ANN) shows consistency of an accuracy rate of stock market prediction. Four methods in predicting stock market had accuracy around 95%. The highest accuracy achieved by using Signal Processing/Gaussian Zero-Phase Filter (GZ-Filter) with near to 98% prediction accuracy. Overall result showed that the ability of an artificial neural network (ANN) in stock market prediction accuracy was satisfactory and achieved high accuracy. Research work of Nonita Sharma et al. [8] emphases on the prediction of future stock market index values based on historical data. The experimental evaluation was based on historical data of 10 years of two indices, namely, CNX Nifty and S&P Bombay Stock Exchange (BSE) Sensex from Indian stock markets. The prediction performance of the proposed model was compared with that of well-known Support Vector Regression. Technical indicators were selected as inputs to each of the prediction models. The closing value of the stock price was the predicted variable. Results showed that the proposed scheme outperforms Support Vector Regression and can be applied successfully for building predictive models for stock prices prediction. In their work the focus was to predict the future values of stock market indices based on the previous stock values using regression. Experiments were carried out on ten years of historical data (January 2006 to December 2016) of two indices namely CNX Nifty and S&P BSE Sensex from Indian stock markets. The predictions were made for 1–10, 15, 30, and 40 days in advance.

Proposed Methodology

Here we tend to area unit planned System that may work with Improved level of recommendation. System are going to be developed with tongue Processing (NLP) of computer science and Convolutional Neural Network (CNN) of Deep Leaning. Natural Language Processing technology can facilitate system to search out companies with excellent news in terms of live performance in market. That may facilitate to create selection of best entertainer in market. NLP will classify news in positive and negative sets and can provide performance graph of selected organization. Supported to that we are going to get sense of best performing company. Natural Language Processing provides to system NLP (Natural Language Processing) that will work on our news for detection merchandise and unhealthy of its impact.

A. Architecture

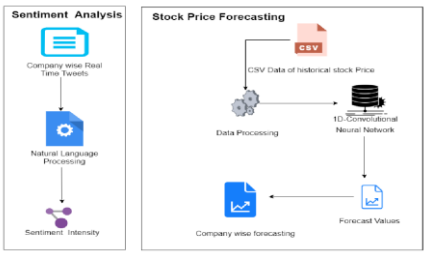

Fig 4.1 Proposed System Architecture

Features of Proposed System

Proposed system has advantage of multiple platforms for input file for model development. System has been developed with over one algorithmic rule thence Prediction guarantees is magnified. Live updates area unit concerned in prediction thence it is often used for live recommendation.

B. Algorithms

First 1D CNN layer:

The primary layer defines a filter (or conjointly known as feature detector) of height ten (also known as kernel size). Solely shaping one filter would enable the neural network to be told one single feature within initial layer. This won't be comfortable, so we will outline N variety of filters. This permits North American country to coach N completely different options on the primary layer of the network. The output of the primary neural network layer is in a very somatic cell matrix type. Every column of the output matrix holds the weights of 1 single filter. With the outlined kernel size and considering the length of the input matrix, every filter can contain variety of weights.

Second 1D CNN layer:

The outcome from the first CNN will be sustained into the second CNN layer. We will again characterize 100 unique channels to be prepared on this level. Following a similar rationale as the primary layer, the yield grid will be of size 62 x 100.

Maxpooling layer:

A pooling layer is frequently utilized after a CNN layer so as to lessen the unpredictability of the yield and forestall over fitting of the information. In our model we picked a size of three. This implies the size of the yield lattice of this layer is just 33% of the info network. Third and fourth 1D CNN layer: Another arrangement of 1D CNN layers follows so as to learn more significant level highlights. The yield lattice after those two layers is a 2 x 160 grid.

Average pooling layer:

One all the more pooling layer to additionally abstain from over fitting. This time not the most extreme worth is taken but rather the normal estimation of two loads inside the neural system. The yield lattice has a size of 1 x 160 neurons. Per include identifier there is just one weight staying in the neural system on this layer.

Dropout layer:

The dropout layer will haphazardly dole out 0 loads to the neurons in the system. Since we picked a pace of 0.5, half of the neurons will get a zero weight. With this activity, the system turns out to be less touchy to respond to littler varieties in the information. Along these lines it should additionally expand our precision on concealed information. The yield of this layer is as yet a 1 x 160 network of neurons. Fully connected layer with SoftMax activation: The last layer will decrease the vector of stature 160 to a vector of six since we need to anticipate ("Jogging", "Sitting", "Strolling", "Standing", "Upstairs", "Ground floor") because we have six classes. This decrease is finished by another lattice augmentation. Softmax is utilized as the initiation work. It powers every one of the six yields of the neural system to summarize to one. The yield worth will hence speak to the likelihood for every one of the six classes.

Sentiment Intensity analyzer

Business: In advancing field firms use it to build up their strategies, to know clients' sentiments towards product or entire, anyway people answer their battles or item dispatches and why customers don't get some product [1] [5].

VADER Sentiment Analysis

VADER content opinion investigation utilizes a human-driven methodology, consolidating substance examination and exact approval by exploitation human raters and hence the information on the gathering.

Five Easy Heuristics

1) Lexical alternatives aren't the sole things inside the sentence that affect the opinion. Their territory unit elective talk segments, similar to accentuation, capitalization, and modifiers that conjointly give feeling. VADER conclusion examination thinks about these by thinking about 5 simple heuristics. The aftereffect of those heuristics zone unit, once more, measured exploitation human raters. For Ex.

[1] I Like that.

[2] I Like that!!!

VADER assumption examination mulls over this by enhancing the slant score of the sentence relative to the quantity of shout focuses and question marks finishing the sentence. VADER first figures the estimation score of the sentence. In the event that the score is certain, VADER includes a specific experimentally acquired sum for every accentuation mark (0.292) and accentuation (0.18). On the off chance that the Score is negative , VADER subtracts.

2) The second heuristic is capitalization.

[1] amazing work.

[2] AMAZING work.

And so VADER takes this under consideration by incrementing or decrementing the sentiment score of the word by zero.733, betting on whether or not the word is positive or negative, severally. 3) The third heuristic is that the use of degree modifiers. view example “effing cute” and “sort of cute”. The result of the modifier within the 1st sentence is to extend the intensity of cute, whereas within the second sentence, it's to decrease the intensity. VADER maintains a booster wordbook that contains a collection of boosters and dampeners. The result of the degree modifier conjointly depends on its distance to the word it’s modifying. Farther words have a comparatively

smaller exacerbating result on the bottom word. One modifier adjacent to the base word adds or subtracts zero.293 to the slant score of the sentence, wagering on whether the base word is sure or not. A second modifier from the base word includes/subtract ninety fifth of zero.293, and a third includes/subtracts ninetieth.

4) The fourth heuristic is that the shift in polarity thanks to "but". In many cases, "but" interfaces 2 provisions with contrastive conclusions. The prevailing opinion, in any case, is that the last one. for example, " I like it, but I don't wish to use that anymore " the essential provision "I like it" is sure, anyway the other VADER actualizes a "but" checker. Fundamentally, all conclusion bearing words before the "but" have their valence decreased to five hundredth of their qualities, though those when the "but" increment to a hundred and fiftieth of their qualities. 5) The fifth heuristic is looking at the tri-gram before a feeling loaded lexical component to get extremity nullification. Here, a tri-gram alludes to an assortment of 3 lexical choices. VADER keeps up a stock of useless words. Refutation is caught by increasing the opinion score of the assessment loaded lexical component by partner degree experimentally decided cost - 0.74.

A. *1D CNN Algorithm:*
The Algorithm of a 1D-CNN is formed through the following important steps: **Input: Dataframe (train_data , test_data)**

Process: Build 1D CNN Model def Model():
define model
add filter (kernel) size to each layer model.add(layers) model.add(kernel size) add pooling layer add dropout value model activation layer fit model with training and testing data
Model summary
Output: Prediction = model.predict(test data)
Accuracy = (accuracy_score(Y_test,Y_pred)*100)

Results and Discussions

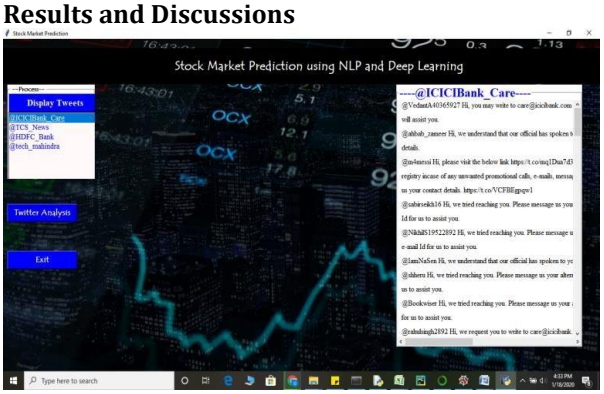

Fig 5.1 Display Tweet window

Fig. 5.1 displays the base window of proposed system, where it contains control panel with buttons which are bind with events to load tweets from twitter according to selected company. Scraped tweets are displayed in canvas window at the right side of base window. Scraped tweets are being displayed in text form in the canvas frame. So here we can say that user will be able to display all live tweets of selected company. So for our project we have taken ICICI bank's official twitter handler's tweets as an example.

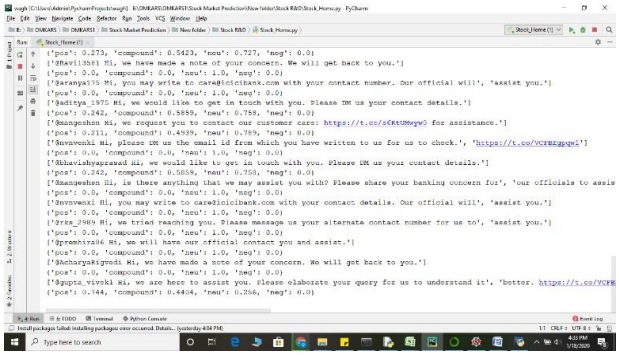

Fig. 5.2 Polarity Result

Fig. 5.2 displays score of polarity on tweets which were scraped from twitter. VADER library that supported our proposed system in terms of getting sentiment of tweets has given the polarity report with Positive , Compound , Neutral and Negative in the form of percentage of each sentiment. So according to polarity score we will get an idea about the positive and negative tweets.

Conclusion

In an early research on stock prediction were totally based on random walks, machine learning ,numerical prediction and support vector machine but with the introduction of behavioral finance, the people's literacy about market were considered while predicted about stock movement. Making it more efficient we used the idea of sentiment analysis of Stock Tweets through NLP technology. We implemented the idea by collecting sentiment data and stock price data and built a CNN model for prediction and forecasting, also in the last we measured the prediction accuracy. Results showed that we have achieved a polarity and based on this polarity we measured top ten well performing companies in given sectors. In future we will attempt to execute more calculations and all the more new methods planning to give live proposal to securities exchange financial specialists. Additionally our emphasis will be on entire securities exchange for forecasting.

References

[1] Batra, Rakhi, and Sher Muhammad Daudpota. "Integrating StockTwits with sentiment analysis for better prediction of stock price movement." In *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pp. 1-5. IEEE, 2018.

[2] Wang, Yaojun, and Yaoqing Wang. "Using social media mining technology to assist in price prediction of stock market." In *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, pp. 1-4. IEEE, 2016.

[3] Sharma, Ashish, Dinesh Bhuriya, and Upendra Singh. "Survey of stock market prediction using machine learning approach." In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, vol. 2, pp. 506-509. IEEE, 2017.

[4] Zhang, Ze, Yongjun Shen, Guidong Zhang, Yongqiang Song, and Yan Zhu. "Short-term prediction for opening price of stock market based on self-adapting variant PSO-Elman neural network." In *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pp. 225228. IEEE, 2017.

[5] Shah, Dev, Haruna Isah, and Farhana Zulkernine. "Predicting the Effects of News Sentiments on the Stock Market." In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 4705-4708. IEEE, 2018.

[6] Peng, Du. "Analysis of Investor Sentiment and Stock Market Volatility Trend Based on Big Data Strategy." In *2019 International Conference on Robots & Intelligent System (ICRIS)*, pp. 269-272. IEEE, 2019.

[7] Firdaus, Muhammad, Swelandiah Endah Pratiwi, Dionysia Kowanda, and Anacostia Kowanda. "Literature review on Artificial Neural Networks Techniques Application for Stock Market Prediction and as Decision Support Tools." In *2018 Third International Conference on Informatics and Computing (ICIC)*, pp. 1-4. IEEE, 2018.

[8] Sharma, Nonita, and Akanksha Juneja. "Combining of random forest estimates using LSboost for stock market index prediction." In *2017 2nd International Conference for Convergence in Technology (I2CT)*, pp. 1199-1202. IEEE, 2017.

[9] Moholkar, Kavita, and Suhas Patil. "Hybrid CNN-LSTM Model for Answer Identification." In 2019 *International Journal of Recent Technology and Engineering (IJRTE)*
ISSN: 2277-3878, Volume-8 Issue-3, September 2019