International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 1087-1092 www.ijaem.net ISSN: 2395-5252

Stock Market Prediction Based On Twitter Sentiment Analysis

Aniruddha Ganthade, Eshani Kulkarni, Anit Agrawal, Sanskar Garodia, Dr. Prashant Lahane

MIT World Peace University
MIT World Peace University
MIT World Peace University
MIT World Peace University
MIT World Peace University

Submitted: 10-05-2022 Revised: 17-05-2022 Accepted: 20-05-2022

ABSTRACT– Stock market prediction is the most emerging topic in computer engineering and the financial sector because new methods and approaches on this matter acquire profit uniformly. In this paper, we used sentiment analysis and machine learning algorithms to establish the relationship between public opinion and the changes in the stock markets. We use tweets tweeted by the various users of a particular company as data to predict the mood of the people and combine the same data with previous days' stock market values to predict the stock market movements. Twitter is one of the most popular social media platforms which provides every user to express their views about a company or occasion or anything to the public. Specifically, we fetch all the live tweets from twitter related to the particular company using the API. Then we preprocessed the tweets. All the special characters, emojis,Ascii characters are removed from the tweets in the data-preprocessing. Then, we performed sentimental analysis on the tweets and segregated them into neutral, negative and positive tweets. For stock market prediction we fetched live stock dataset through yahoo finance API. The stock value dataset of the company along with the tweets tweeted about the same company are taken as input to the machine learning model to predict the stock values of the same company. We used Apple company stock data with 3 ML models named LSTM,ARIMA, Linear Regression. The main aim of the project is to combine the previous stock data of the company with the current tweets related to the company and predict the stock values for the same company.

I. INTRODUCTION

Stock market prediction is the most emerging topic in computer engineering and the financial sector. Due to the huge profits from the stock market in current situations, it has attracted people from the business as well as engineering side. Stock price forecasts have forever been a subject of interest for most financiers and monetary analysts. Nevertheless, judging the best opportunity to buy or advertise has been a very difficult project for financiers cause skilled are additional many determinants that may influence stock prices.

Today folks square measure putting their comments and opinions on social media which might be shared by others additionally. Sentiment classification may well be tired word/phrase level, sentence level and document level. Sentiment analysis has currently become the dominant approach used for extracting sentiment and appraisals from on-line sources. Judgment analysis focuses on dividing language units into 2 categories: objective and subjective, whereas sentiment analysis tries to divide the language units into 3 categories; negative, positive, and neutral.

Twitter is a social media platform where around several tweets square measure sent daily. Newspaper headlines additionally give info associated with the exchange which might even be used for prediction functions. exploitation the twitter information, a prediction method will be performed. varied tweets associated with completely different firm square measure obtained within the Twitter API.There could also be several tweets that aren't used for prediction functions. Live twitter information will be extracted from the twitter API and analyzed exploitation of the classifier. Stock information will be fetched from the Yahoo finance API to investigate the worth. varied machine learning algorithms square measure accustomed to train the model to predict the stock worth.

We mix machine learning and public sentiment. This has been achieved by employing a hybrid formula that uses sentiment analysis and LSTM to predict ensuing day stock values and also the public’s sentiment, that helps North American countries to correlate the market conditions and public sentiment. publically offered Twitter information is employed to perform sentiment analysis and yahoo finance to induce stock values.

The economical Market Hypothesis (EMH) states that exchange costs square measure for the most part driven by new info and follow a stochastic process pattern. Although this hypothesis is widely accepted by the analysis community as a central paradigm governing the markets generally, many folks have tried to extract patterns within the manner stock markets behave and answer external stimuli.

Stock market prediction is the method of evaluating the longer term worth of the stock of a selected company, therefore giving a thought of gain or loss to the investors to speculate in this explicit company stock. Social media plays a very important role in predicting the stock worth

II. SYSTEM DESIGN

We used the stock market values of the company through yahoo finance api. Parallely, the tweets are fetched through tweepy. Then we used the twitter data for sentiment analysis of the tweets we gave the moods of the people like positive, neutral and negative while tweeting as an output. Then the above moods are processed with the stock values of the company to predict the future stock values of the company. The learnt model similarly because the previous stock exchange prices and mood values ar employed by the portfolio management system that runs the model to predict the long run value and uses the expected values to form acceptable buy/sell choices.

The below figure shows the flow diagram of our project.

Fig. 1 Architecture of System

III. DATASET

In this project, we used two main datasets-

- The Stock market data for the company was fetched using Yahoo Finance api which includes the close, open, low, adjusted and high values for a given day.
- Publicly available Tweets more than 476 million, corresponding to more than 17 million users were fetched through tweepy. The tweets included the username, text and timestamp for every tweet.

Data Preprocessing

The data obtained from the sources mentioned above had to be pre-processed to form it appropriate for reliable analysis. we have a tendency to pre-processed knowledge within the following manner-

While the Twitter knowledge was offered for all days lying within the giving amount, values obtained exploitation Yahoo! Finance was absent for weekends and different holidays once the market was closed. so as to finish this knowledge, we have a tendency to approximate the missing values employing a plano concave performance. So, if the worth on a given day is x and also the next offered datum is y with n days missing in between, we have a tendency to approximate the missing knowledge by estimating the primary day when x to be (y+x)/2 then following constant methodology recursively until all gaps square measure stuffed. This approximation is even because the stock knowledge typically follows a plano concave performance, unless in fact at anomaly points of explosive rise and fall. Then we have a tendency to solely took the worth of the last a pair of years solely.

Tokenization: every tweet is split into individual words referred to as tokens. This method is completed to interrupt the text, separated by whitespace characters. Removal of stop words:

Words like “the”,“a”, “an”, “she”, “he”, “on”, “by”, etc don't seem to be needed for sentiment analysis. These square measures are referred to as stop words, which are removed before the sentiment analysis method. Regex Matching: Special characters like “URL”, “!”, “#”, “@” square measure all removed and replaced by whitespaces.

IV. SENTIMENT ANALYSIS

Twitter is a popular social networking and microblogging service that allows people to tweet their feelings, views or opinions about a particular occasion or topic. Earlier people could tweet about 140 characters in length but nowadays length is 280 characters . Due to the character of this microblogging service (quick and short messages), folks use acronyms, create writing system mistakes, use emoticons and different characters that have specific special meanings.

Twitter tweets analysis is a one of the most important parts of our project since the output of the twitter sentiment analysis ws combined to stock data of the company to predict the stock values of the given company for next seven days. We consider the tweet tweeted by the user as either positive or negative about the company. If the overall polarity of the tweet is positive the stock price of the company will increase or vice versa.

Tweet Cleaning: To make the tweet easy to read we initially removed certain characters from the tweets. We removed emojis from the tweets. Later we removed hindi characters from the tweets. Also the tweets fetched from twitter & were written as &amp so we rewrite &amp to &.

Target: Twitter data fetched contained symbol “@” which refer to other users mentioned in the tweets. Referring people or organization helps to know what the tweet is referring to

Hashtags: Users use hashtags to refer to particular topics. This helps users to spread their thoughts more efficiently. We acquire Twitter data (tweets) from Twitter using the python library and Twitter’s developer portal. 9oo;

V. MARTINGALE STOCK TRADING STRATEGY

With prediction markets growing in variety and in prominence in numerous domains, the development of a modeling framework for the behavior of costs on listed contracts has become Associate in Nursing a more and more necessary endeavor. The Martingale strategy suggests doubling up on losing bets and reducing winning bets by half. It is essentially a strategy that not only promotes a loss-averse mentality that tries to improve the odds of breaking even, but also increases the chances of severe and quick losses.

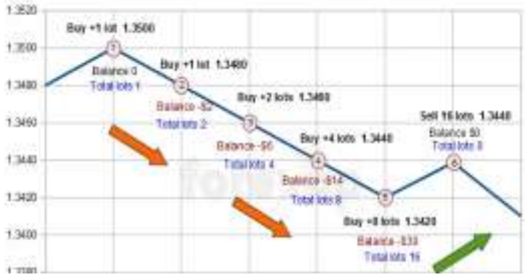

Fig. 2 The Martingale Strategy

Some cons of Martingale trading strategy are -

- The amount spent on commercialism will reach immense proportions once simply some transactions.
- If the monger runs out of funds and exits the trade whereas victimization the strategy, the losses baby-faced is calamitous.
- There is an opportunity that the stocks stop commercialism at some purpose in time.
- The risk-to-reward magnitude relation of the Martingale Strategy isn't cheap. whereas victimization the strategy, higher amounts are spent with each loss till a win, and therefore the final profit is barely adequate the initial bet size. The strategy ignores dealings prices related to each trade.
- There area unit limits placed by exchanges on trade size. Therefore, a monger doesn't receive Associate in Nursing an infinite variety of possibilities to double a bet.

VI. MODEL LEARNING AND PREDICTION

LSTM

Long Short Term Memory or LSTM may be a kind of repeated neural network that's capable of learning order dependencies in sequence prediction issues. It's feedback connections that facilitate it to method sequences of information. It conjointly has internal state cells that perform as long or short term memory cells. The output is regulated by these state cells. This property is incredibly helpful once we ought to rely upon previous inputs instead of the newest ones. As time passes it's less doubtless that the output is passionate about terribly previous inputs, these inputs are a unit forgotten by it through their forget gates, the forget gates is simply an increasing issue of zero.9, that's among twelve steps the issue becomes zero.9^12 ≈ 0.282. Equations utilized by forget, input and output gates in LSTM facilitate LSTM to contemplate previous inputs in predicting succeeding output and conjointly the forget gate equation helps in forgetting terribly previous inputs and therefore the equation is :

$f_t = \sigma(w_f\,(h_t{-}1, x_t) + b_f\,)$
$i_t = \sigma(w_i(h_t{-}1, x_t) + b_i)$
$o_t = \sigma(w_0(h_t{-}1, x_t) + b_0)$

We separated data in 8:2 ratio 80% as training data and 20% to test data.In the proposed algorithm we use LSTM to predict closing values of stock for next seven days. The primary step is to scrape the information from yahoo finance.The data obtained from yahoo finance is then normalized. the information is then scaled so as to suit the values between 0(minimum) and 1(maximum) also know minmax normalization

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Fig. 3 Formula for Rescaling

Result: We used LSTM to predict the stock price of the AAPL company. The information utilized in this study was obtained from Twitter and yahoo finance. we have a tendency to collect all two knowledge from the past 10 years i.e 2012 to 2022. for every day, the gap, lowest shutting and highest values of the stock were obtained. When coaching the information we have a tendency to take a look at our model on the remaining. we have a tendency to then prepare a chart so as to visualize the distinction between our take a look at and trained prediction values. The below chart shows the comparison between the expected and also the actual values of stock worth for Apple company.

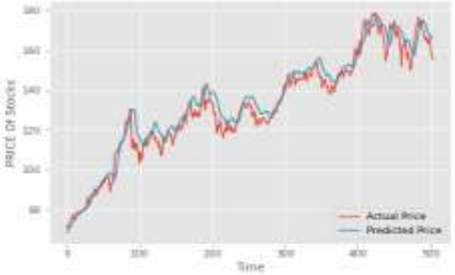

Fig 4. Result of LSTM Algorithm for APPLE Company

Arima

An ARIMA model is a class of mathematical models for resolving and forecasting occasion succession dossiers. It definitely caters to a suite of standard makeup later succession dossiers, and as such specifies a natural still powerful form for making able period succession forecasts. ARIMA stands for Autoregressive Joined Affecting Average.This puzzle is explanatory, capturing the key facets of the model itself. Concisely, they are:

- AR: Autoregression. Is a model which uses the dependent relationship between an observation and some number of lagged observations.
- I: Integrated. The use of differencing of raw observations in order to make the time series stationary.
- MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Adopting AN ARIMA model for a statistic assumes that the underlying method that generated the observations is AN ARIMA method. This could appear obvious, however it helps to encourage the necessity to verify the assumptions of the model within the raw observations and within the residual errors of forecasts from the model. ARIMA is nominative by 3 ordered parameters (p,d,q). Where:

- p is the number of time lags
- d is the number of times in which data have or had past values been subtracted
- q is the order of moving average models. For building an ARIMA model, it is necessary to make sure that the data is stationary.

Results: We used ARIMA to predict the stock value of AAPL company and the result we got was pretty accurate as we can see all predicted and actual value are almost same

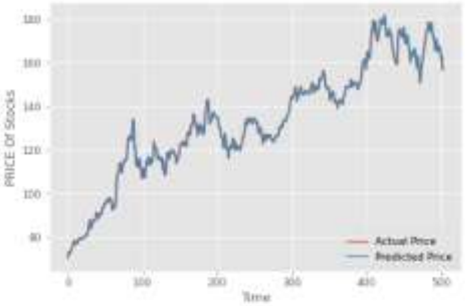

Fig 5. Result of ARIMA Algorithm for APPLE Company

Linear Regression

Linear Regression may be a machine learning formula that is predicated on supervised learning. It performs a regression task. Regression models a target prediction worth supported freelance variables. it's principally used for locating out the connection between variables and prediction. disagreement|completely different} regression models differ in support – the type of relationship between dependent and freelance variables they're considering and also the range of freelance variables being employed. regression toward the mean performs the task to predict a variable quantity worth supporting a given experimental variable; thus, this regression technique finds out a linear relationship between input and output. Hence, the name is regression toward the mean.The regression toward the mean model provides a sloping line representing the connection between the variables. think about the below image:

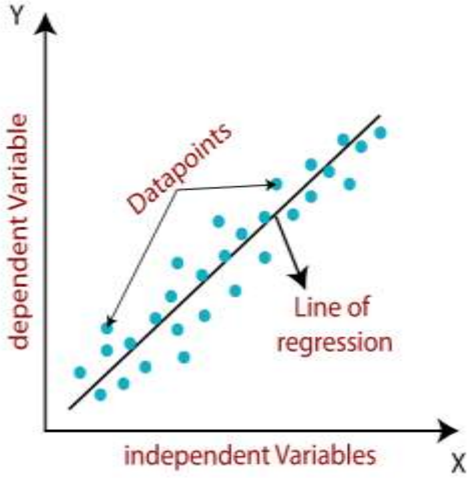

Fig 6 Graph of Linear Regression

Hypothesis function for Linear Regression :

$$y = \theta_1 + \theta_2.x$$

Where,

X is independent variable
Y is dependent variable

When operating with regression toward the mean, our main goal is to search out the simplest match line meaning the error between foreseen values and actual values ought to be decreased . The simplest match line can have the smallest amount of error.

The different values for weights or the constant of lines (θ1, θ2) offers a distinct line of regression, thus we'd like to calculate the simplest values for a0 and a1 to search out the simplest match line, thus to calculate this we tend to use price perform.

Result: We used Linear Regression to predict the stock value of AAPL company and the result we got is shown in the comparative graph shown below.

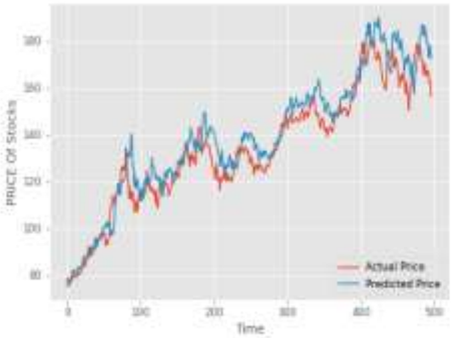

Fig 7. Result of Linear Regression Algorithm for APPLE Company

VII. CONCLUSION

Unlike the traditional stock exchange prediction systems, our novel approach combines the emotions of people through the twitter tweets and predicts the longer term values of the stock exchange. The twitter feeds are obtained for a listed company and sentiment polarity of the tweets are calculated for the prediction of stock news, whether or not it's positive, negative or neutral. The stock numbers are foreseen mistreatment metric capacity unit models Finally, a mix of sentiment polarity points and also the foreseen costs offer associate degree economical results to the stock exchange forecasters once to shop for or sell their stocks. The Martingale commercialism Strategy offers an associate degree outlook of commercialism the stock at foreseen costs and displays the probable profit/loss.

Our future work focuses on working with RSS feed stock news and multi language tweets which improves the accuracy of the system. Also we intend to fetch data or social contents from other social platforms other than twitter for effective prediction.

REFERENCES

[1]. Anshul Mittal and Arpit Goel. Stock Prediction Using Twitter Sentiment Analysis.

[2]. Padmanayana, & Varsha, & K, Bhavya. (2021). Stock Market Prediction Using Twitter Sentiment Analysis. International Journal of Scientific Research in Science

and Technology. 265-270. 10.32628/CSEIT217475.

[3]. Mittal, A. and Goel, A., 2012. Stock prediction using twitter sentiment analysis. Stanford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf), 15, p.2352.

[4]. Bharathi, S. and Geetha, A., 2017. Sentiment analysis for effective stock market prediction. International Journal of Intelligent Engineering and Systems, 10(3), pp.146-154.

[5]. Karlemstr and, R. and Leckström, E., 2021. Using Twitter Attribute Information to Predict Stock Prices. arXiv preprint arXiv:2105.01402.

[6]. Nti, I.K., Adekoya, A.F. and Weyori, B.A., 2020. A systematic review of fundamental and technical analysis of stock market predictions. Artificial Intelligence Review, 53(4), pp.3007-3057.

[7]. Klein, D., Prediction Market Prices as Martingales: Theory and Analysis

[8]. Panday, Harsh, et al. "Stock Prediction using Sentiment analysis and Long Short Term Memory." European Journal of Molecular & Clinical Medicine 7.2 (2020): 5060-5069.

[9]. Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." Proceedings of the workshop on language in social media (LSM 2011). 2011.

[10]. Zhang, Lei, Shuai Wang, and Bing Liu. "Deep learning for sentiment analysis: A survey." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8.4 (2018): e1253.