iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.43132

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Analysis of Text Data for Stock Prediction

Rakshit Vastrad[1], Akul G Devali[2], Rohit G Urs[3], Nithin D[4]
[1, 2, 3, 4]*Department of Computer Science and Engineering, Vidyavardhaka college of Engineering Mysuru, India,*

Abstract: Accounting for price fluctuations and understanding people's emotions can help to improve stock price forecasting. Only a few models can decipher financial jargon and have stock price change datasets that have been labelled. In this project, we used text mining techniques to extract high-quality data from news and tweets published by legitimate businesses on the internet, allowing us to analyse, decide, and update our database for future use. In this paper, we propose an information gathering and processing framework that combines a natural language processing tool with our algorithms. We use natural language processing and machine learning techniques to make predictions. The result demonstrates the algorithm's ability to foresee favorable outcomes.

I. INTRODUCTION

The stock market has grown to be one of the most important components of the economy not only in developed countries, but also in developing and third-world countries. Almost all large scale and mid-level companies have their stocks listed on the stock market. Added to this, along with traditional factors such as infrastructure, number of employees, expansion over the years etc., the health of the company stock and the profit it has made to investors has become a trademark factor in the assessment of the overall growth and development of any country. These trends are extensively followed and have begun to have increasingly bigger impacts on the lives of common people. The economic recession of 1929 and the real estate market crash of 2008 which cost millions of people their jobs and livelihoods is a startling proof of this fact. Hence it is the need of the hour to predict stock movements which will enable us to maximise our benefits, understand individual risk apatite and make safe decisions while investing in the stock market.

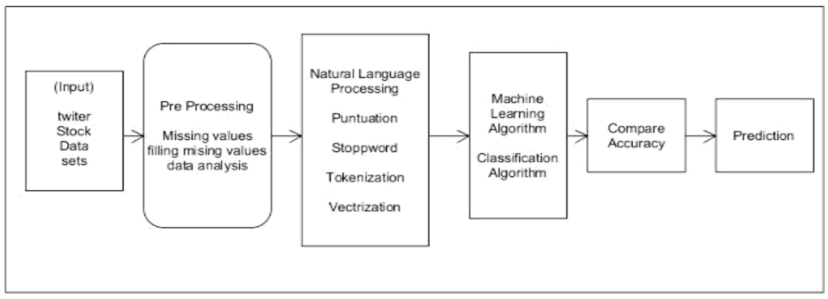
The numerous factors that influence each decision we make it difficult to make a stock market decision. As a result, in order to make the best stock market move, extensive analysis is required, which may include price trend, market nature, company stability, various stock news and so on. While studies and methods developed earlier were adequate for analysing the financial charts and patterns we see a new trend emerging in this decade. The movement of stock prices is increasingly being determined by traditional factors but also the talk and rumours about these stocks. This can be attributed to the decentralisation of the media and it's division into orthodox sources such as newspapers, television channels on one hand and various organisations on the other, many of which exist online only. Many people get their news about the world around them, which invariably include stocks from social media posts and trends on popular apps such as Twitter and Instagram.

The goal of this research is to extract fundamental data from relevant news sources and analyse and, in some cases, forecast the stock market from the standpoint of the average investor. In order to provide a more complete and comprehensive perspective of the market to the user we have developed a system that caters not only to the financial charts but also the chatter and sentiments of people regarding a particular stock. Based on a review of existing business text mining research, we developed a framework that combines our text parser and analyser algorithm with a natural language processing tool open source, which converts texts to numeric data with the help of labelling in order to analyse, retrieve, and forecast stock market investment decisions from any text data source using machine learning and text mining.

II. METHODOLOGY

1) Collecting of datasets from tweeter.
2) Data pre Processing: Analysis of collected data and identifying the missing values, filling missing values and selection of attributes
3) The preprocessed data is passed as input to the NLP to convert text data to structure to numeric data
4) The NLP data is passed to machine learning algorithms for prediction
5) Regression algorithms is used for prediction
6) Build the model file using trained data and compare the algorithm result

©IJRASET: All Rights are Reserved



A. Naïve Bayes

It is a supervised machine learning algorithm based on the Bayes theorem and assumes that feature pairs are independent. It is presumptively based on the assumption that none of the variables in the dataset are correlated with one another, but rather Nave. It is the most user-friendly, and it works well for large datasets and datasets with categorical data. Furthermore, when compared to other traditional models, the training time for this model is extremely short.

B. Random Forest Classifier

It is an ensemble learning method based on multiple individual decision trees which are created at training time. This means it takes individual tree prediction results into account when determining the outcome, resulting in improved performance. This is superior to the prediction result of any single tree. For tasks which involve classification, the Random Forest Classifier returns the class which has been selected by most of the trees as the result. However in the tasks which involve regression, the mean of the classes is found, which is then returned as the result.

C. Gradient Boosting Repressor

Another Ensemble technique (boosting) is to create predictors in a sequential rather than independent manner. Each predictor learns from the mistakes of the previous predictor. As a result, obtaining accurate predictions takes less time and iterations.

D. Logistic Regression

When there is a dependent (target) variable, a statistical model is used. To begin, linear regression is used to fit the data. Then, for predicting the probabilities of various classes of data, a logistic function is used. A sigmoid function is used to convert these probabilities to binary form, which aids in making actual predictions. In a low-dimensional dataset, it is less prone to overfitting the model.

E. XGBoost

eXtreme Gradient Boosting is a machine learning algorithm that is supervised. It is built with decision trees and employs a gradient boosting framework. It is also effective with tabular or structured data. Because of system optimization and algorithmic enhancements, XGBoost provides good performance.

III. LITERATURE REVIEW

1) Authors Rakhi Batra and Sher Muhammad Daudpota used a model whose objective is to predict next day's stock movement. NLP, Support Vector Machine (SVM) analysis shows that market data and people opinion have a positive correlation. Their proposed work has an accuracy of 76.65% in stock prediction. Large datasets yield low accuracy.
2) In this work the objective is to predict investor sentiment from a StockTwits Platform, the methodology used is Bayesian Networks. The results provide a better understanding of the predicted relationships among sentiments and their related feature.
3) The Authors Sunil Kumar Khatri, Ayush Srivastava have performed sentimental analysis on the data extracted from Twitter and StockTwits. The polarity index along with market data is supplied to an artificial neural network to predict the result. Data samples can't be for a longer period.
4) The Authors have designed and implemented machine learning models to forecast stock prices using StockTwits using Machine Learning and Linear Regression with a accuracy of 65%.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

5) In this work carried out by the authors the main objective id detecting the actual financial opinions behind texts. Neural Networks, Deep Learning is used to identify the financial opinions from texts. The Real time analysis can't be performed in this particular model
6) Joseph coelho et al have developed models based on linear regression and neural networks.In which they have used the year worth stocktwit data to analyse stock prices but the drawback of these models is that it performs poorly on high dimensional data
7) In this work the authors have tried to predict stock prices by combining news and social sensing with financial statements using Machine Learning, the outcome of the analysis is investment signals for buying or selling stocks. The drawback of this paper is that the specific implementation has not been tested in terms of scalability.
8) The objective of this work is to identify trends in machine learning research using text mining methods. A text mining approach is applied in this work is to detect trends in terms in research articles published over time. The authors examined 21906 papers published in six top machine learning journals between 1988 and 2017. The analysis shows how various terms used in machine learning research have evolved over the past three decades. The study will guide the next generation of researchers to the significant research areas of machine learning.
9) In this work the author presents a overview of text mining using various machine learning techniques. This work has also highlighted the importance of automating the text mining process and it's challenges as the human capabilities are far more effective in the current scenario. It is also a significant development that sentiment analysis is used to predict the outcome of elections at both stage and national level. However with the help of a machine learning techniques, language processing and visualization, it is possible to design and develop an extraordinary mining system.
10) Mukul Jaggi et al have made a Fin-ALBERT model to predict the stock price by text mining of stocktwits for 25 companies including FAANG. The authors have used some Ml algorithms to train model like Naïve Bayes, Random forest classifier, XG Boost etc. The drawback of this model is that the best algorithm yields only 59% accuracy.
11) The authors have used some old models like classification model using decision tree to predict buy or sell the stocks. The historical data of stocks for two years has been used to train the model. They have analyzed and trained the data of three companies (ARBK,MECE,UAIC).But the few drawbacks of this paper is that it has only analyzed data of three companies, there are better models like Machine learning ,neural networks and many and if the stock price changes due to factors, this model is unable to anticipate investors' influence, political events, the economy, and general economic conditions.
12) In deep learning, long short-term memory (LSTM) is a type of recurrent neural network (RNN), which has feedback connections. LSTM can process both individual data points (like images) as well as entire sequences of observations (such as speech or video inputs).

No.	Algorithms	Result
[1]	Machine Learning, NLP, Support Vector Machine (SVM)	accuracy 76.65% .
[2]	Bayesian Networks	Highest Accuracy : 72.74 % Lowest Accuracy : 58.21 %
[3]	Artificial Neural Network	Senitimental Score : [highest] 0.85, [lowest] 0.39
[4]	Machine learning, Linear Regression	Accuracy 65%

[5]	Neural Networks, Deep Learning	Accuracy : highest 90.7%, lowest : 71.3%.
[6]	Machine learning, Linear Regression	Accuracy level : highest
[7]	Machine Learning (ML)	Investing signals include recommendations for buying or selling stocks based on the results of the analysis.
[8]	Machine Learning	Study analysis allows upcoming researchers to gain insight into a significant area of research that involves machine learning.
[9]	Data Mining, Natural Language Processing	This study presents an overview of text mining approach with its techniques, tools and applications.
[10]	Fin ALBERT	Accuracy : 59%
[11]	ID 3, C4.5	The highest accuracy obtained is 54.9 %
[12]	LSTM	Training RMSE : 0.00983 Testing RMSE : 0.00859.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

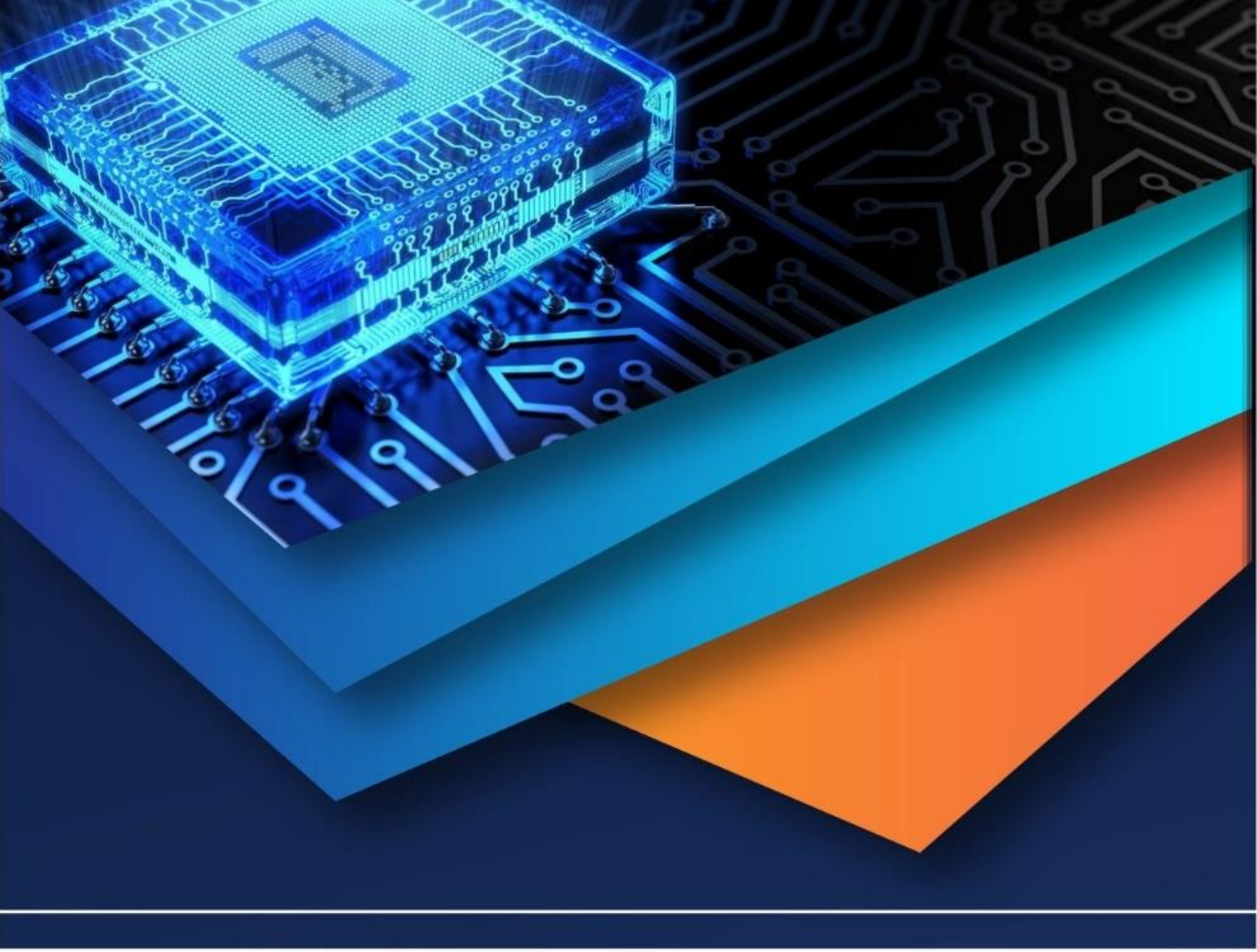
IV. CONCLUSION

Finally, rather than sentiments, to categorize the messages, the change in stock prices was used as a labelling technique. Currently, only a few experiments on this labelling technique are being carried out. Comparing the Percentage change technique with two labels to the other labelling techniques tested, this technique produced the best results in all models. The FinALBERT model is affected by hyperparameter settings and dataset size. Despite being pre-trained on a much smaller dataset, the model performed well when compared to the other models. For FinALBERT models to perform well, a large dataset with numerous training steps must be pre-trained, which was not possible due to hardware constraints. When compared to traditional models, the training time for the transformer-based model was excessively long.

REFERENCES

[1] Rakhi Batra Department of Computer Science Sukkur IBA University rakhi.bhatra@iba-suk.edu.pk Sher Muhammad Daudpota Department of Computer Science Sukkur IBA University sher@iba-suk.edu.pk Integrating StockTwits with Sentiment Analysis for better Prediction of Stock Price Movement.
[2] Alya Al Nasseri , Allan Tucker, and Sergio de Cesare Big Data Analysis of StockTwits to Predict Sentiments in the Stock Market.
[3] Sunil Kumar Khatri, Ayush Srivastava Amity Institute of Information Technology Amity University Uttar Pradesh, Noida, India sunilkkhatri@gmail.com, skkhatri@amity.edu ayush.idea77@gmail.com Using Sentimental Analysis in Prediction of Stock Market Investment.
[4] Scott Coyne, Praveen Madiraju and Joseph Coelho Department of Mathematics, Statistics and Computer Science Marquette University Milwaukee, WI, USA Forecasting Stock Prices using Social Media Analysis.
[5] Liang Zhang, Keli Xiao, Hengshu Zhu, Chuanren Liu, Jingyuan Yang, Bo Jin CADEN: A Context-Aware Deep Embedding Network for Financial Opinions Mining.
[6] Joseph Coelho, dawson d'Almeida, Scott Coyne, Nathan Gilkerson, katelyn Mills, Praveen Madiraju Social Media and Forecasting Stock Price Change.
[7] Traianos-Ioannis Theodorou, Alexdros Zamichos, Michalis Skoumperdis, Anna Kougioumtzidou, kalliopi Tsolaki, Dimitris Papadopoulos, Thanasis Patsios, George Papanikalaou, Athanasios Konstantinids, Anastasios Drosou and Dimitrios Tzovaras An AI-Enabled Stock Prediction Platform Combining News and Social Sensing with Financial Statements.
[8] Deepak Sharma^1, Bijendra Kumar^1, Satish Chand^2 1 Department of Computer Science Engineering, Netaji Subhash Institute of Technology, New Delhi, India 2. School of Computer And Systems Sciences, Jawaharlal Nehru University, New Delhi, India Trend Analysis in Machine Learning Research Using Text Mining.
[9] Abhishek Kaushik and Sudhanshu Naithani Kiel university of Applied Science Kurukshetra University A Comprehensive Study of Text Mining Approach.
[10] Mukul Jaggi *, Priyanka Mandal, Shreya Narang, Usman Naseem and Matloob Khushi Text Mining of Stocktwits Data for Predicting Stock Prices.
[11] Qasem A. Al-Radaideh, Adel Abu Assaf, Eman Alnagi, qasemr@yu.edu.jo,abuassaf@gmail.com,ealnagi@philadelphi a.edu.jo Predicting stock prices using data mining techniques.
[12] Murtaza Roondiwala, Harshal Patel, Shraddha Varma Predicting Stock Prices Using LSTM.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)