iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** X **Month of publication:** October 2021

DOI: https://doi.org/10.22214/ijraset.2021.38662

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: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

A Systematic Study on Sentiment Analysis based on Text Mining and Deep learning for Predictions in Stock Market Trends Through Social and News Media Data

Suvigya Jain[1], Mubin Modi[2], Divyanshu Gupta[3]
[1]*Medicaps University*

Abstract: ***Stock Market has always been one of the most active fields of research, many companies and organizations have focused their research in trying to find better ways to predict market trends. The stock market has been the instrument to measure the performance of a company and many have tried to develop methods that reduce risk for the investors. Since, the implementation of concepts like Deep Learning and Natural Language Processing has been made possible due to modern computing there has been a revolution in forecasting market trends. Also, the democratization of knowledge related to companies made possible due to the internet has provided the stake holders a means to learn about assets they choose to invest in through news media and social media also stock trading has become easier due to apps like robin hood etc. Every company now a days has some kind of social media presence or is usually reported by news media. This presence can lead to the growth of the companies by creating positive sentiment and also many losses by creating negative sentiments due to some public events. Our goal in this paper is to study the influence of news media and social media on market trends using sentiment analysis.***
Keywords: ***Deep Learning, Natural Language Processing, Stock Market, Sentiment analysis.***

I. INTRODUCTION

Formerly,Time series forecasting in stock market analysis are among the most difficult tasks inside the field of data science, and as a consequence, numerous academics had attempted to resolve this issue. As a result, a lot of effort has gone into developing model for stock price prediction. Machine learning, big data, deep learning, and sentiment analysis employing natural language processing have all been used to address issues.

A mix of pattern recognition, deep learning, advanced analytics, and natural language processing has showed promise in anticipating market movements for short-term trading.There are quite a variety of methods that can be used to get the required outcomes and improve the computation accuracy.

The utilization of data sets for sentiment analysis, on the other hand, significantly enhanced the outcome. Also it's crucial to pick the right source of data for sentiment analysis. Using multiple news interfaces, we may leverage media stories and internet news services from several well-known news organizations such as CNN, The Verge, and others.

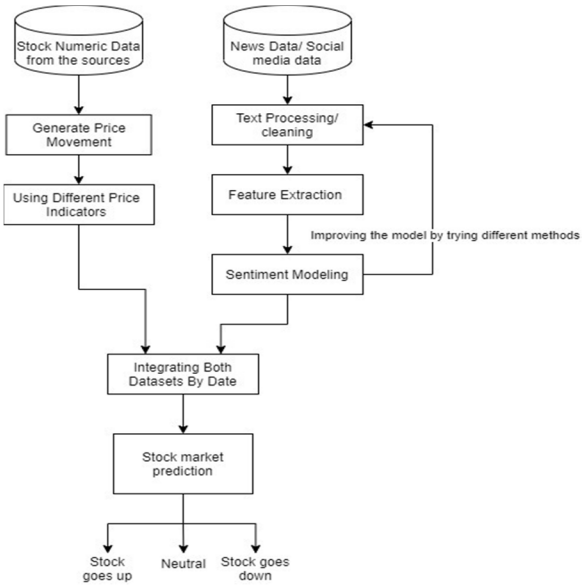
We may also utilize data from social media platforms like Twitter and Reddit to forecast market movements. In this article, we'll look at how news and social media data can be used to forecast stock market patterns and make predictions out of it, by which Buyers will be able to turn a profit if they can reliably assess stock value volatility.

Forecasting how well the financial sector will move is among the most difficult tasks as there are many variables at play, such as lending rates, geopolitics, and economic expansion, that make the stock market turbulent and difficult to foresee precisely.

II. METHODOLOGY

Since Stock market prediction is well deep analyzed topic many researchers have proposed different methods for the task , However, all of them adhere to the broad structure depicted in the diagram below. To begin, the user must obtain the essential stock data from sites such as yahoo finance,Metastock,Morning star,Bloomberg.com,Alpha Vantage etc.They use variety of indicators like On-Balance Volume, Accumulation/Distribution Line, Average Directional Index, Relative Strength Index., Stochastic Oscillator etc. for analytical purposes.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com



Firstly, accumulation the textual data related to the target organization is done for further sentiment analysis for identification of objective to be validate. This data can be gathered from News articles, Internet news or social media platforms like Twitter,Facebook,Instagram etc. depending on the requirements using APIs like Rapid API, Twitter API and many other valid sources.Another way to gather data is to Web-scraping text data from the internet using Python libraries like beautiful-soup. It is important to note that for this method to work the target must have a social presence on the internet so that we can perform analysis through knowing sentiment about the target from their allocated users.

After the data is gathered it has to be processed and cleaned for further processes. For that removal of redundant data entry is must and extrapolation of the missing values to determine nearby figures. For the text data we need to remove the irrelevant texts and only retain what is useful, data that gives us the sentiment about the target. There are various methods to extract features from the text data using techniques like Word Embedding, Word2Vec, GloVe, T-SNE and many more methods as per the requirement of the analysis that has to be done. Using the extracted features we will create a sentiment model for our analysis. We will then merge the stock data and Text data by using their respective dates. All these data points will then be used to model our stock prediction model. We can uses various deep learning models like CNNs(Conventional Neural Networks), RNNs(Recurrent Neural Networks) and LSTM(Long Short-Term Memory) in combination with the mentioned feature extraction methods to create a good stock prediction model.

For example-Have purposed an approach for Stock Market Prediction Using LSTM (Long Short-Term Memory)and Sentiment Analysis using historical stock data from 1st January 2014 to 31st December 2018 along with text data from news headlines regarding the stock. All of this data is fed into an LSTM (Long Short-Term Memory) network to get a prediction for the month of January 2019. This approach is compared to a naive approach of only using LSTM(Long Short-Term Memory) without sentiment analysis of news headlines. They aim to show the correlation of news headlines in influencing the fluctuations in stock prices.

Furthermore, social media data for the same task which have proposed a novel filtered based classification model on real-time stock market technical data and its social medial comments to predict the trend of each stock for the classification problem. Features are extracted from the comments and technical data for stock trend prediction. Finally, each stock trend is predicted using the proposed non-linear SVM classifier in different levels of time frames. To get to our point details like feature extraction and the algorithms used for different methods of prediction might change but all the generalized process will be the carried out in same procedure.

III. CASE-STUDIES FOR SENTIMENT ANALYSIS IN GENERAL

It is important for user to consider their source of data in their modelling process. There two popular source of data are News and social media data. Both have their pros and cons. News data is generated by industry professionals, its well thought out and it is reliable most of the times. It very likely to accurately correlate with market trends. We can study the effect of news data using the following case studies.

The study effect of combining technical indicators from stock prices based on more than five years of Hong Kong Stock Exchange data using four different sentiment dictionaries and news sentiments from textual news articles, and make prediction models be able to learn time series sequential information in an intelligent way. They have built a stock prediction system and proposed an approach that converts historical prices into technical indicators that summarize aspects of the price information, and models news sentiments by using different sentiment dictionaries and represents textual news articles by sentiment vectors, then constructs a two-layer LSTM neural network to learn the sequential information within market snapshots series, then constructs a fully connected neural network to make stock predictions. The proposed approach outperforms the baselines in both validation and test sets using two different evaluation metrics. The LSTM incorporating both information sources outperforms models that only use single one information source, in both individual stock level and sector level. Among the four sentiment dictionaries, finance domain-specific sentiment dictionary (Loughran–McDonald Financial Dictionary) models the new sentiments better, which brings more prediction performance improvements (at most 120%) than the other three dictionaries (at most 50%).

In case study of predicting movement of stock market,they have used Dow Jones Industrial Average (or Dow Index) that tracks and combines performance on stock markets for 30 large, publicly-owned companies trading on the US stock market to study the effect of news media on stock market. They used the historical data from years 2008 – 2020 along with headlines from the top-5 highest rated news companies to predict the movement of the Dow index. They also compare the shallow Machine Learning (ML) and Deep Leaning (DL) approaches to predict the Dow Index movement based on the news headlines. They use of two different text representation methods as an addition to all the approaches. They have taken two data sets: one that includes dates before the COVID-19 pandemic and the other that includes the dates during the COVID-19 pandemic up to the June 17th 2020 date, to study how their models handle unprecedented events due to pandemic. Results showed that these the Deep Learning models were more effective than the Machine learning models. They also demonstrated how useful news data is to predict stock market trends. They prove that their model is capable of predicting the economy stability due to the unrepresented events such as COVID-19.

Through the following case studies, we can establish that news data is very useful in predicting market trends, we to also consider that news data can sometimes be faked to try to influence the market however, the there very little evidence to support this idea and it is a small part of the larger available news dataset.

Case study on Social media data: Social media data is raw reaction of different individuals with diverse views of a particular topic.It is very subjective it may or may not also correlate with the actual market situations. There are different ways in which social media can impact market trends, it can be directly impact because of certain chain events or it can passively impact market trends because of organizations marketing strategies, it depends on the public perception of the company. We will study the following effects using some case studies.

We can consider the case of GameStop; GameStop attracted the worldwide financial community's attention when its price increased from $16 to $347 within a span one month during January 2021. However, a unique feature of this upsurge was the contest between a large number of small retail investors and institutional investors such as hedge funds rather than the GameStop underlying fundamentals. Thus, this episode provides a unique case of the impact of investor sentiments on share prices. The episode started when the Wallstreet bets group on the Reddit platform, comprising a large number of small retail investors, started its campaign against the big hedge funds in order to show the power of crowd against the big hedge funds, involved with GameStop. This group of the retail investors on the Reddit platform (the Reddit investors) countered the massive short positions on the GameStop shares, contracted by the involved hedge funds, by taking long positions in the shares of GameStop. The basic reason behind the massive short selling was the weak fundamentals of the GameStop company from last three years, and there was an expectation of a further decline in share prices. However, despite the weak fundamentals, the Reddit investors started buying GameStop shares that led to a 21-fold increase within one month inflicting huge on the institutional investors. For example, one of the biggest US hedge funds, Melvin Capital Management, lost 53 percent of its investments in January 2021 due to the short squeeze on the GameStop shares. According to media-wise Reddit investor sentiments may have positively affected the GameStop returns during the recent episode of tension between the social media investors and the involved hedge funds. It implies that investors can get reasonable returns in short run following big social media groups like Reddit.

Another case study that examines the impact of sentiments expressed in social media and news media on market returns. Data used to establish results has been collected from an online discussion forum, HotCopper (HC), and investigated the impact of users' sentiments expressed on this forum on the stocks listed in ASX 50. News articles related to these stocks were collected from Google Finance. Messages and articles are classified using Bernoulli Naïve Bayes classifier and sentiments are analyzed using the bullishness index. Our study shows a significant effect of sentiments from social media and news media on market returns. We find that sentiments from social media have a long lasting and stronger effect on market returns than the sentiments from news media.

IV. TOOLS AND PACKAGES USED FOR ANALYSING SENTIMENTS

The emoticons are the most commonly used instruments for determining the polarity (favourable and unfavourable effect) of a communication. Emoticons are based on facial expressions.Despite the fact that there are several different types of flowers or faces , they all represent sad or happy emotions.Other tool is indeed the Multilingual Research and Number Of words count, which uses a vocabulary of words with their classed divisions to analyse not just to favourable and unfavourable but also psychological, mental, and shaping of a text.These software uses an index approach to grade a sentiment on a supervised approach where a chosen sentiment determine the feelings of a person and make classification of the user depending on labeled data-sets which produces outputs based on relative inputs.SentiStrength is also another utility that calls "one of most prominent built sentiment classification software." It assigns ratings to favourable and unfavourable terms in literature using a lexicon - based approach. Several essential predictors are suggested for determining the sensation pole. It is detailed inside the Senti Word-net tool-set. Senti Word-net is a semantic library for text categorization and sentiment distributed data mining that is accessible to the public. This is built on Word-net , an American lexical thesaurus that groups descriptors, plurals, actions, and other words into sequences.

V. STRENGTHS AND WEAKNESSES OF SENTIMENT ANALYSIS

Every day, a huge proportion of consumer media is produced on social networking websites, which is a prominent feature to analyze the sentiment of different people around the globe.This data can then be used to make greater risk - based decisions, enhance organization effectiveness, provide improved products/services, as, eventually, enhance consumer lifestyles and social connection in order to establish a healthier future. The influence of measuring sentiments on goods, services, and activities, for instance, provides theory suggests that managers to own facts and criteria to make decisions. A further instance is planning commission supervisors, who may have the chance to improve government operations and identify environmental stewardship challenges quite proficiently based on what they know about sentiments of peoples.

Processing is an area has drawbacks because it is hard to execute due to the complexities of natural speech as well as the features of the textual comments. The assessment of tweets is an aspect of this, since they are frequently accompanied by keywords, symbols, and urls, making deciphering the conveyed emotion challenging. Furthermore, autonomous procedures are required, which necessitate voluminous data of labeled posts or vocabulary databases with moral terms correlated with sentiment polarity. Another key feature would be that assessments are appropriate for Pronunciation, which is a restriction for those other dialects.These will affect the accuracy of sentiments for labeled data sets and will make difficult to analyze the situation precisely.

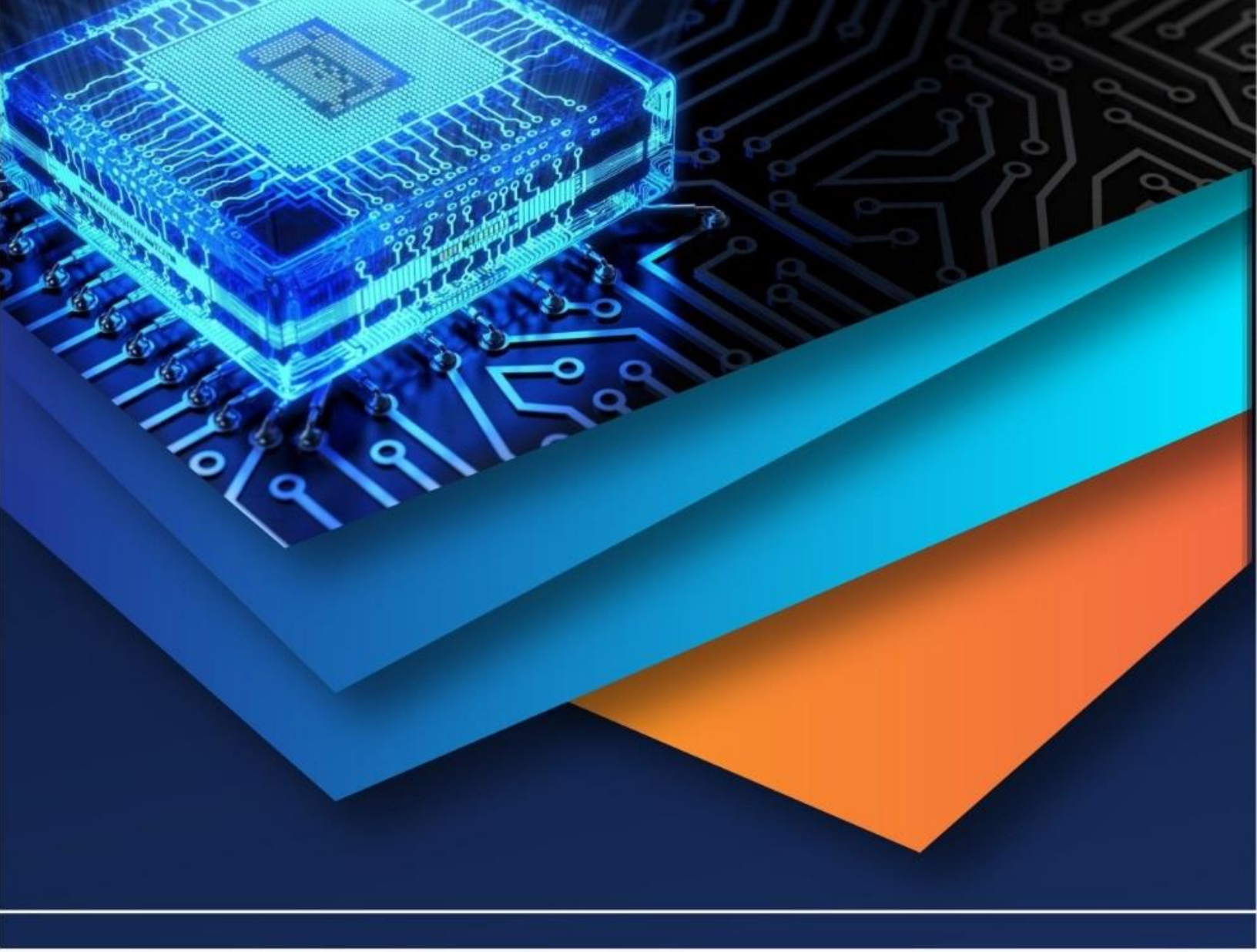
VI. CONCLUSIONS

This research uncovered a wide range of methodologies, procedures, tools and packages, and including strengths and weakness used for the approaches to foresee prediction and trends of stock market using natural language processing and deep learning.When analyzing trend analysis, it is necessary, according to some resources, to incorporate a wide range of pattern methodologies. In order to properly enable share proof decision making by comprehending News and social media data models, these models must also adhere to high research rules. Models must be evaluated against well-defined selection criteria and then validated both internally and externally before being used to drive future predictions. Just a few researches have reported the level of evidence required to analyze data-sets and provide in making future decisions. Also, there are various analytical software,tools and packages which are further used for sentiment predictions based on data that brings accuracy and precision in making insights which helps in further analysis to train and gain knowledge to form an integrating market structure based on research analysis of stocks through their historical data in making precise decisions.

REFERENCES

[1] Z. Wang, S. B. Ho, and Z. Lin, “Stock market prediction analysis by incorporating social and news opinion and sentiment,” IEEE Int. Conf. Data Min. Work. ICDMW, vol. 2018-November, pp. 1375–1380, 2019, doi: 10.1109/ICDMW.2018.00195.
[2] N. Strauss and C. H. Smith, “Buying on rumors: how financial news flows affect the share price of Tesla,” Corp. Commun., vol. 24, no. 4, pp. 593–607, 2019, doi: 10.1108/CCIJ-09-2018-0091.
[3] L. P. Maguluri and R. Rengaswamy, “An efficient stock market trend prediction using the real-time stock technical data and stock social media data,” Int. J. Intell. Eng. Syst., vol. 13, no. 4, pp. 316–332, 2020, doi: 10.22266/IJIES2020.0831.28.
[4] Y. Ge, J. Qiu, Z. Liu, W. Gu, and L. Xu, “Beyond negative and positive: Exploring the effects of emotions in social media during the stock market crash,” Inf. Process. Manag., vol. 57, no. 4, 2020, doi: 10.1016/j.ipm.2020.102218.
[5] W. Chen, C. K. Yeo, C. T. Lau, and B. S. Lee, “Leveraging social media news to predict stock index movement using RNN-boost,” Data Knowl. Eng., vol. 118, pp. 14–24, 2018, doi: 10.1016/j.datak.2018.08.003.
[6] M. R. Vargas, B. S. L. P. De Lima, and A. G. Evsukoff, “Deep learning for stock market prediction from financial news articles,” 2017 IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl. CIVEMSA 2017 - Proc., pp. 60–65, 2017, doi: 10.1109/CIVEMSA.2017.7995302.

[7] Y. Duan, L. Liu, and Z. Wang, "COVID-19 Sentiment and the Chinese Stock Market: Evidence from the Official News Media and Sina Weibo," Res. Int. Bus. Financ., vol. 58, p. 101432, 2021, doi: 10.1016/j.ribaf.2021.101432.
[8] J. Clarke, H. Chen, D. Du, and Y. J. Hu, "Fake news, investor attention, and market reaction," Inf. Syst. Res., vol. 32, no. 1, pp. 35–52, 2021, doi: 10.1287/isre.2019.0910.
[9] C. Guan, W. Liu, and J. Y. C. Cheng, "Using Social Media to Predict the Stock Market Crash and Rebound amid the Pandemic: The Digital 'Haves' and 'Have-mores,'" Ann. Data Sci., no. 0123456789, 2021, doi: 10.1007/s40745-021-00353-w.
[10] J. Y. Huang and J. H. Liu, "Using social media mining technology to improve stock price forecast accuracy," J. Forecast., vol. 39, no. 1, pp. 104–116, 2020, doi: 10.1002/for.2616.
[11] M. Kesavan, J. Karthiraman, T. Ebenezer Rajadurai, and S. Adhithyan, "Stock Market Prediction with Historical Time Series Data and Sentimental Analysis of Social Media Data," Proc. Int. Conf. Intell. Comput. Control Syst. ICICCS 2020, no. Iciccs, pp. 477–482, 2020, doi: 10.1109/ICICCS48265.2020.9121121.
[12] X. Li, P. Wu, and W. Wang, "Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong," Inf. Process. Manag., vol. 57, no. 5, p. 102212, 2020, doi: 10.1016/j.ipm.2020.102212.
[13] G. Jariwala, H. Agarwal, and V. Jadhav, "Market," pp. 1–5, 2020.
[14] S. Kalra and J. S. Prasad, "Efficacy of News Sentiment for Stock Market Prediction," Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019, pp. 491–496, 2019, doi: 10.1109/COMITCon.2019.8862265.
[15] A. Sarkar, A. K. Sahoo, S. Sah, and C. Pradhan, "LSTMSA: A Novel Approach for Stock Market Prediction Using LSTM and Sentiment Analysis," 2020 Int. Conf. Comput. Sci. Eng. Appl. ICCSEA 2020, no. i, pp. 4–9, 2020, doi: 10.1109/ICCSEA49143.2020.9132928.
[16] Y. Liu, J. Trajkovic, H. G. H. Yeh, and W. Zhang, "Machine Learning for Predicting Stock Market Movement using News Headlines," 2020 IEEE Green Energy Smart Syst. Conf. IGESSC 2020, 2020, doi: 10.1109/IGESSC50231.2020.9285163.
[17] K. Prachyachuwong and P. Vateekul, "Stock trend prediction using deep learning approach on technical indicator and industrial specific information," Inf., vol. 12, no. 6, 2021, doi: 10.3390/info12060250.
[18] D. Shah, H. Isah, and F. Zulkernine, "Predicting the Effects of News Sentiments on the Stock Market," Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018, pp. 4705–4708, 2019, doi: 10.1109/BigData.2018.8621884.
[19] L. Nemes and A. Kiss, "Prediction of stock values changes using sentiment analysis of stock news headlines," J. Inf. Telecommun., vol. 5, no. 3, pp. 375–394, 2021, doi: 10.1080/24751839.2021.1874252.
[20] N. Ivanova, V. Gugleva, M. Dobreva, I. Pehlivanov, S. Stefanov, and V. Andonova, "We are IntechOpen , the world ' s leading publisher of Open Access books Built by scientists , for scientists TOP 1 %," Intech, vol. i, no. tourism, p. 13, 2016.
[21] C. G. Schmidt, D. A. Wuttke, G. P. Ball, and H. S. Heese, "Does social media elevate supply chain importance? An empirical examination of supply chain glitches, Twitter reactions, and stock market returns," J. Oper. Manag., vol. 66, no. 6, pp. 646–669, 2020, doi: 10.1002/joom.1087.
[22] S. Bharathi, A. Geetha, and R. Sathiynarayanan, "Sentiment analysis of twitter and RSS news feeds and its impact on stock market prediction," Int. J. Intell. Eng. Syst., vol. 10, no. 6, pp. 68–77, 2017, doi: 10.22266/ijies2017.1231.08.
[23] P. Jiao and A. Walther, "Social Media, News Media and the Stock Market," SSRN Electron. J., no. January 2016, 2016, doi: 10.2139/ssrn.2755933.
[24] P. Statement, "Stock Prediction Using Twitter Sentiment Analysis Problem Statement."
[25] T. Hu and A. Tripathi, "The Effect of Social and News Media Sentiments on Financial Markets Regular Paper," pp. 1–16.
[26] F. Jin, W. Wang, P. Chakraborty, N. Self, F. Chen, and N. Ramakrishnan, "Tracking multiple social media for stock market event prediction," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10357 LNAI, pp. 16–30, 2017, doi: 10.1007/978-3-319-62701-4_2.
[27] Z. Umar, M. Gubareva, I. Yousaf, and S. Ali, "A tale of company fundamentals vs sentiment driven pricing: The case of GameStop," J. Behav. Exp. Financ., vol. 30, p. 100501, 2021, doi: 10.1016/j.jbef.2021.100501
[28] S. Usmani and J. A. Shamsi, "News Headlines Categorization Scheme for Unlabelled Data," 2020 Int. Conf. Emerg. Trends Smart Technol. ICETST 2020, 2020, doi: 10.1109/ICETST49965.2020.9080726.
[29] D. Valle-Cruz, V. Fernandez-Cortez, A. López-Chau, and R. Sandoval-Almazán, "Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis During Pandemics: A Comparative Study of the H1N1 and the COVID-19 Periods," Cognit. Comput., no. 0123456789, 2021, doi: 10.1007/s12559-021-09819-8.
[30] X. Zhang, J. Shi, D. Wang, and B. Fang, "Exploiting investors social network for stock prediction in China's market," J. Comput. Sci., vol. 28, pp. 294–303, 2018, doi: 10.1016/j.jocs.2017.10.013.
[31] Y. Zhang and H. Liu, "Stock market reactions to social media: Evidence from WeChat recommendations," Phys. A Stat. Mech. its Appl., vol. 562, p. 125357, 2021, doi: 10.1016/j.physa.2020.125357.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY