INVESTIGATE SENTIMENTAL DATA FROM TWITTER FOR BUSINESS STRATEGIES USING MACHINE LEARNING

[1]Ms. P. BUSHRA ANJUM, [2]Mr. D. KOTESWARA RAO, [3]B. YAMINI, [4]T. VISHNU VARDHAN, [5]P. SRI SOWMYA, [6]SK. IRFAN

[1]Assistant Professor, Department of Data Science, NRI Institute of Technology, Visadala Road, Perecherla, Andhra Pradesh, India

[2]Associate Professor & HOD, Department of Computer Science-Data Science, NRI Institute of Technology, Visadala Road, Perecherla, Andhra Pradesh, India

[3,4,5,6]B.Tech Scholars, Department of Computer Science-Data Science, NRI Institute of Technology, Visadala Road, Perecherla, Andhra Pradesh, India

Abstract: Businesses are increasingly seeking sentiment analysis research to quickly gauge consumer perceptions of their brand, enabling them to make informed adjustments to their business strategies. This research utilizes modeling techniques and a dataset of 70,000 tweets to evaluate sentiment polarity. By leveraging Natural Language Processing (NLP) and word embeddings, the tweets are effectively trained and classified, yielding positive results. It was observed that removing stop words from the data could decrease the precision of sentiment classification. The research implements a comprehensive pre-processing stage to enhance the tweets readability for standard NLP techniques. Each dataset example includes two tweets and their corresponding sentiment, facilitating the application of supervised machine learning.

The study proposes sentiment analysis models based on Random Forest, logistic regression, and Support Vector Machines (SVM). The primary aim is to improve the analysis of emotions, categorizing tweets into positive or negative sentiments. Among the methods, the Random Forest technique achieves the highest accuracy, reaching up to 95%.

Keywords: Natural Language Processing (NLP), Machine Learning, Random Forest, Logistic Regression, Support Vector Machines (SVM).

I. INTRODUCTION

This paper will examine the entire development process, provide a thorough specification, talk about the background information, and evaluate how the paper was carried out. The report will also examine the project's technological evolution, outlining important phases and provide analysis of choices made as well as results from testing any algorithmic models developed during the process.

Twitter has emerged as a prominent platform for discussion of intense emotions, making it a valuable source of information for analyzing sentiments. Sentiment analysis is the technique of examining text to detect its underlying emotional tone. With the rise of social media platforms like Twitter, analysis of sentiment has become an essential tool for businesses, associations, and governments seeking to comprehend public opinion and form well-informed perspectives. Natural Language Processing (NLP) methods are extensively employed for sentiment analysis as they enable machines to comprehend and interpret human language. NLP techniques can analyze tweets in real time, identify the sentiment conveyed in tweets, and provide insights into prevailing trends and patterns in public sentiment. Machine learning algorithms, which fall under the umbrella of NLP, can acquire knowledge from vast datasets and accurately predict the sentiment

of new tweets. In this investigation, we aim to assess the efficacy of ML systems in conducting sentiment analysis on Twitter using NLP methodologies.

To classify tweets as favorable, negative, or neutral., we will utilize a dataset that includes tweets from the opening day of the "FIFA World Cup 2022", held in Qatar. This dataset encompasses information such as the date of creation, number of likes, tweet source, tweet content, and sentiment. We will preprocess this data to eliminate any noise and subsequently use machine learning methods such as Vader, XG Boost, Random Forest, and LSTM (Long Short-Term Memory). To improve accuracy, we use count vectorizers and genism in our models. Machines cannot interpret letters or words. When dealing with text data, we must represent it numerically so that the machine can interpret it. Count vectorizer is a method for translating text to numerical data. Gensim is an open-source Python package for NLP. The Gensim package this allows us to create word embeddings by instruction word2vec classifiers on a particular corpus using either the CBOW or skip-gram approaches [1].

A few years ago, machine learning algorithms have been applied to twitter sentiment analysis, allowing researchers to extract valuable insights into public opinion on various topics. A variety of industries, including political campaigns, managing brand reputation, and analyzing customer feedback, can benefit from Twitter sentiment research. The ability to analyze tweets in real-time and monitor public sentiment has become increasingly important in a world where social media has become the primary source of news and information. Our research's first stage is to compile a dataset of tweets about a particular subject. For example, we may collect tweets related to a particular brand, political figure, or social issue. Once we have collected the dataset, we will preprocess the data to remove noise and irrelevant information. This involves removing links, hash tags, and other non-textual data that could affect the accuracy of the sentiment analysis. Next, in order to extract features from the text, we shall employ NLP approaches [2].

II. LITERATURE SURVEY

S. Singh, K. Kumar and B. Kumar et. al. Sentiment analysis technique plays an important role in natural language processing to analyze complex human statements. In the last few years, this technique has become a powerful tool for several social media communication mediums such as WhatsApp, Twitter, Facebook, Instagram, YouTube, LinkedIn, Blog, etc. This paper proposes a machine learning (ML) based method to analyze social media data for sentiment analysis on text data. The presented method is divided into three distinct stages. In the first stage, pre-processing is performed to filter and refine the text data. In the second stage, the feature extraction is performed using the Term Frequency and Inverse Document Frequency (TF-IDF) technique. Moreover, during the third stage, the extracted features are supplied to make predictions for the classifier. The experiments are carried out on a publicly available Twitter dataset for

US Airlines. Several ML techniques are utilized for analysis and classification. The results are reported for different evaluation metrics like accuracy, precision, recall, and F1 score. Finally, the support vector machine yielded the most relevant results [3].

R. Wagh and P. Punde et. al. Social networking sites like twitter have millions of people share their thoughts day by day as tweets. As tweet is characteristic short and basic way of expression. So in this review paper we focused on sentiment analysis of Twitter data. The Sentiment Analysis sees as area of text data mining and NLP. The research of sentiment analysis of Twitter data can be performed in different aspects. This paper shows sentiment analysis types and techniques used to perform extraction of sentiment from tweets. In this survey paper, we have taken comparative study of different techniques and approaches of sentiment analysis having twitter as a data [4].

J. Ferdoshi, S. D. Salsabil, E. R. Rhythm, M. H. K. Mehedi and A. A. Rasel et. al. Social media plays a vital role in our daily lives. To understand and interpret emotions and opinions expressed on social media platforms, analyzing sentiment is very important. Our study is based on Twitter sentiment analysis. Our aim is to classify tweets automatically as positive, negative, or neutral based on their content using natural language processing and machine learning algorithms. The dataset we used for our analysis is extracted from the website called mendeley data and also we have added some tweets manually which covers various topics. To remove noise, including URLs, hashtags, punctuations, and user mentions, and to retain essential textual content and emojis, we pre-processed the dataset. Additionally, for our research, we used VADER (Valence Aware Dictionary and sentiment Reasoner) and Transformers-RoBERTa to analyze the sentiment of various tweets. We evaluate the performance of these two models using evaluation metrics such as accuracy, precision, recall and F1-score, and also confusion metrics on the testing set. We also discuss the study's limitations and conclude that machine learning-based sentiment analysis models are a reliable tool for the sentiment analysis of the twitter dataset [5].

M. Khurana, A. Gulati and S. Singh et. al. Text mining is the way toward investigating and breaking down a lot of unstructured content information that can distinguish ideas, designs, subjects, catchphrases and different qualities in the information. Twitter is one of those forums that allow people across the world to put and exchange their views and ideas on several major and minor issues which are revolving around the world every day. Microblogging on twitter gains the interest of data researchers as there is an immense scope of mining and analysing the huge amount of unstructured data in several ways. In this paper, various algorithms for analysing the sentiments of the tweets have been discussed. Further, the performance of these algorithms has been compared based on certain metrics. Certain challenges while doing the study have also been described in terms of improvement and future scope. Since the machine learning algorithms have been performed on an unexplored dataset,

language barriers to these algorithms have also been identified in terms of future scope and current feasibility of the algorithms. The analysis has been performed using classification algorithms - Naïve Bayes, Support Vector Machine and Random Forest. This experimental work has been executed in python and excel has been used to further evaluate and plot some of the results. Since the sentiment of the tweets cannot be beknown, test set has been manually prepared in order to prevent any errors in evaluating accuracy and precision of the models [6].

K. Agarwal, S. Deepa, R. V. SivaBalan and C. Balakrishnan et. al. With the exponential growth of social networking sites, people are using these platforms to express their sentiments on everyday issues. Collection and analysis of people's reactions to purchases of products, public services, etc. are important from a marketing and innovation perspective. Sentiment analysis also called opinion mining or emotion extraction is the classification of emotions in text. This technique has been widely used over the years to determine sentiment within given text data. Twitter is a social media platform primarily used by people to express their feelings about specific events. In this paper, collected tweets about National Education Policy which has been a hot topic for a while; and analyzed them using various machine learning algorithms such as Random Forest classifier, Logistic Regression, SVM, Decision Tree, XGBoost, Naive Bayes. This study shows that the Decision tree algorithm is performing best, compare to all the other algorithms [7].

III. METHODOLOGY

The goal of artificial intelligence (AI) research is to show how machines may be intelligent in contrast to humans and other animals. The goal of artificial intelligence (AI) is to take what humans have been doing for thousands of years, comprehend how we think, and demonstrate how we can utilize this understanding to influence a world that is considerably bigger and more complex than we first realized. Though research in this field started after World War II, the phrase artificial intelligence was first used in 1956. (Norvig and Russell, 2016) In his work "Computing Machinery and Intelligence," published in October 1950, Alan Turing presented one of the earliest approaches to artificial intelligence (AI) aimed at creating robots capable of human-like behavior. Turing suggested a test known as the "Imitation Game," which was eventually dubbed the Turing Test. It involved three players: an interrogator (C), a woman (B), and a male (A). This test was designed to provide an answer to the question, "Can machines think?" The goal of the game was to see if the interrogator could tell which person was a woman and which was a guy based just on how they answered questions. The interrogator gave them the names X and Y. The interrogator was unable to determine the identity of the respondent merely by listening to the tone of voice, for example, as the questions were type written.

The most common usage for machine translation, which is an automatic text translation process, is to convert one language into another. Warren Weaver first proposed it in his article in 1949. "More than

just pointing out the obvious—that linguistic diversity seriously hinders cross-cultural communication and deters international understanding—needs to be said." Making sure we get a word's underlying meaning is a challenge in translating. Sometimes it's impossible to determine whether a word actually means what it says on the page. For instance, if we try to read words through an opaque mask that has just a little hole in it, we won't be able to see the word next to the one we are reading. It's difficult to determine which meaning a word like "fast" conveys—"rapid" or "motionless". It was at this point in the 1960s that the necessity of machine translation began to be discussed. In order to address the necessity of machine translation at the time, a group of scientists known as ALPAC (Automatic Language Processing Advisory Committee) published a paper titled "Languages and Machines" in the early 1960s. The study came to the conclusion that human translators were more efficient and less expensive than machine translation. It was evident by the middle of the 1960s that machine translation was not working, as stated succinctly in the report "There is no emergency in the field of translation." The issue is not trying to solve a hypothetical demand with a hypothetical machine translation.

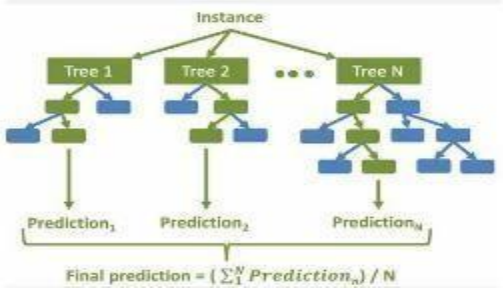

Fig. 1: RF Prediction

The random forest algorithm is applicable to applications involving both classification and regression. The collection of tree-structured classifiers is known as the RF classifier. It is a more sophisticated kind of bagging that incorporates randomization. RF splits each node using the best among a subset of predictors randomly selected at that node, as opposed to utilizing the best split among all variables. From the original data set, a new training data set is generated using replacement. The next step is to use random feature selection to grow a tree. Arboreal trees are not trimmed. With this tactic, RF achieves unprecedented accuracy. Additionally, RF is incredibly quick, resistant to over fitting, and allows the user to create as many trees as they desire. Create bootstrap samples for tree using the original data. Grow an unprimed classification or regression tree for each of the bootstrap samples, but change it so that at each node, randomly sample try of the predictors and select the best split among those variables, instead of selecting the best split among all of the predictors.

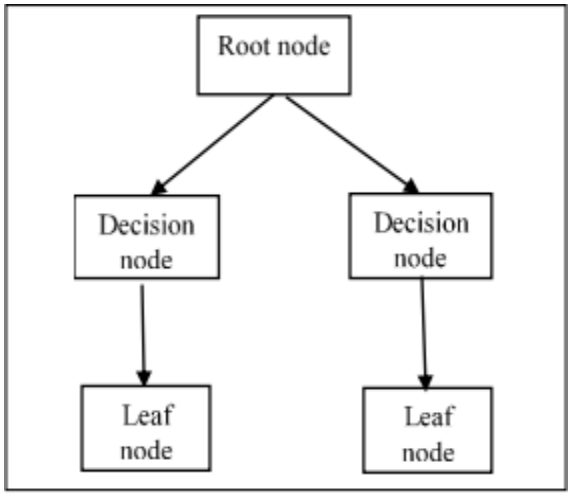

Fig. 2: Structure of DT

The decision tree algorithm is an induction method for data mining that divides a set of records recursively using either a breadth-first or depth-first greedy approach until all of the data items are members of a specific class. Internal, leaf, and root nodes make up the structure of a decision tree. Unknown data records are classified using the tree structure. Impurity metrics are used to determine the appropriate split at each internal node of the tree. The class labels that the data items have been grouped under make up the leaves of the tree. The two stages of the decision tree classification technique are tree building and tree pruning. The process of creating a tree is top-down.

The tree is recursively divided at this stage until every data item is associated with a single class label. Because the training data set is repeatedly accessed, it is extremely laborious and computationally demanding. Trees are pruned from the bottom up. It is employed to reduce over- fitting, or noise or excessive detail in the training data set, in order to increase the algorithm's prediction and classification accuracy. Misclassification error in decision tree algorithms is caused by over-fitting. As the training data set is only scanned once, tree pruning requires less work than the tree growth phase. The decision tree classification in the suggested system gives the user a better way to categories tweets into positive and negative categories. The process involves comparing the maximum frequent things produced by the rules in the training data with the maximum frequent items in the test data. This allows for an easy categorization to be established.

IV. RESULTS

Fig. 3: Test Case 1

Fig. 4: Test Case 2

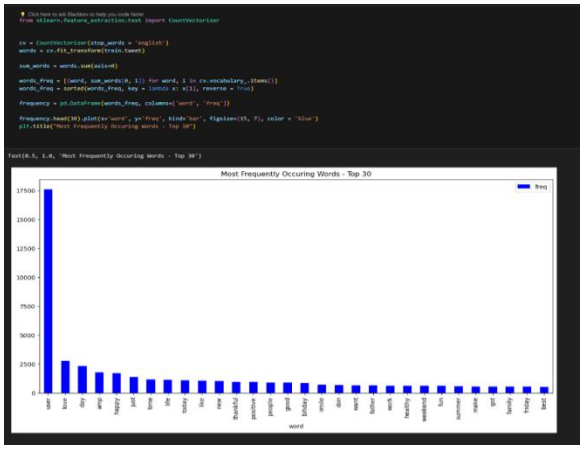

Fig. 5: Testing repeated words in the datasets

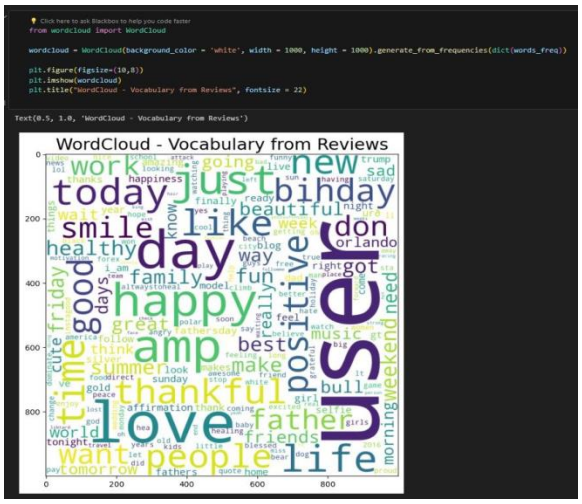

Fig. 6: Using different colors for different words in the dataset

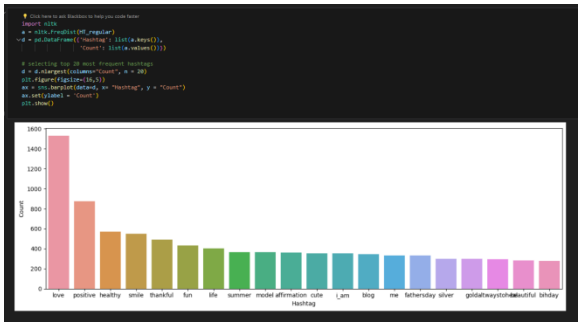

Fig. 7: Testing the percent repeated words in the data set

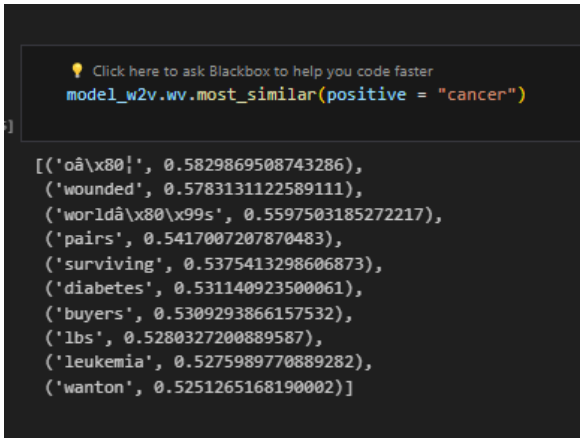

Fig. 8: Testing the results for the dataset (Positive, Negative, Neutral)

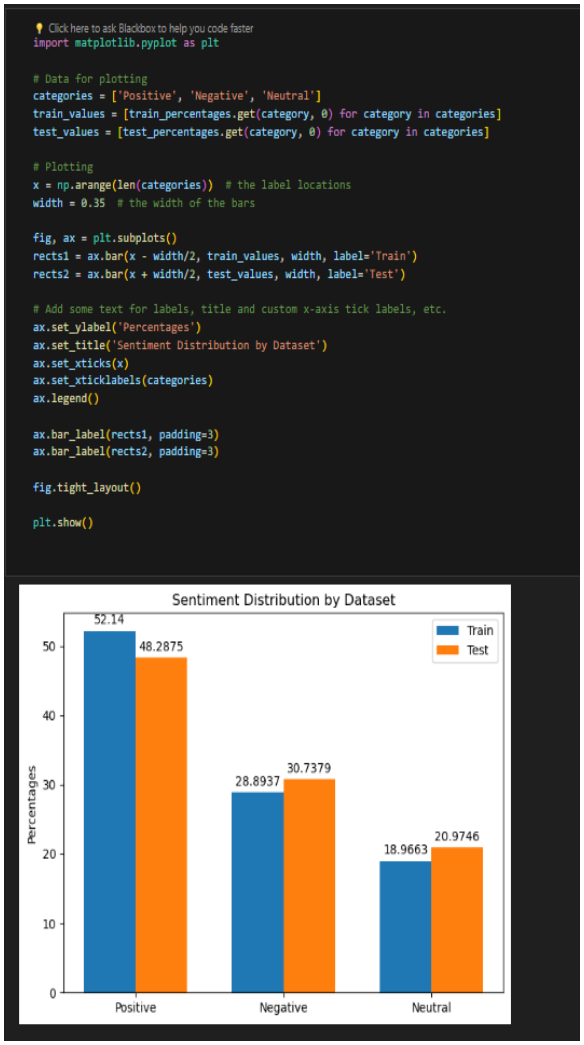

Fig. 9: Difference between the train and test dataset results

V. CONCLUSION

Sentiment analysis, also known as opinion mining, is a field that explores individuals' attitudes, feelings, and sentiments towards specific subjects. This study addresses the primary challenge of sentiment polarity categorization within sentiment analysis. The dataset, sourced from Kaggle.com, includes tweets related to six airlines. Using the Random Forest technique, the study achieved an accuracy of approximately 95% in sentiment analysis. To potentially

improve these results, it is recommended to experiment with alternative machine learning techniques such as logistic regression and support vector machines (SVM). Twitter Sentiment Analysis (TSA) is an emerging area of research focused on determining and evaluating user attitudes and perspectives. The proposed approach in this study involves two phases: tweet classification and pre-processing. Pre-processing involves cleaning the Twitter data by removing unnecessary emojis and handling missing values. The Decision Tree (DT) algorithm is then applied to the pre-processed data for sentiment analysis. The effectiveness of the proposed method is demonstrated using simulated Sanders Twitter data, showing superior classification rates compared to earlier approaches. The method outperforms existing TSA systems in terms of accuracy, F-measure, precision, and recall, improving classification precision and recall rates by approximately 3–15%.

VI. REFERENCES

[1] Khan, Mantasha & Srivastava, Ankita. (2024). Sentiment Analysis of Twitter Data Using Machine Learning Techniques. International Journal of Engineering and Management Research. 14. 196-203. 10.5281/zenodo.10791485.

[2] Richa Dhanta, Hardwik Sharma, Vivek Kumar, Hari Om Singh. Twitter sentimental analysis using machine learning. Int J Commun Inf Technol 2023;4(1):71-83. doi: 10.33545/2707661X.2023.v4.i1a.63

[3] S. Singh, K. Kumar and B. Kumar, "Sentiment Analysis of Twitter Data Using TF-IDF and Machine Learning Techniques," 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India, 2022, pp. 252-255, doi: 10.1109/ COM-IT CON54601.2022.9850477.

[4] R. Wagh and P. Punde, "Survey on Sentiment Analysis using Twitter Dataset," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 208-211, doi: 10.1109/ICECA.2018.8474783.

[5] J. Ferdoshi, S. D. Salsabil, E. R. Rhythm, M. H. K. Mehedi and A. A. Rasel, "Unveiling Twitter Sentiments: Analyzing Emotions and Opinions through Sentiment Analysis on Twitter Dataset," 2023 Computer Applications & Technological Solutions (CATS), Mubarak Al-Abdullah, Kuwait, 2023, pp. 1-7, doi: 10.1109/CATS58046.2023.10424206.

[6] M. Khurana, A. Gulati and S. Singh, "Sentiment Analysis Framework of Twitter Data Using Classification," 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, India, 2018, pp. 459-464, doi: 10.1109/PDGC.2018.8745748.

[7] K. Agarwal, S. Deepa, R. V. SivaBalan and C. Balakrishnan, "Performance Analysis of Various Machine Learning Classification Models Using Twitter Data: National Education Policy," 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India,

2023, pp. 862-870, doi: 10.1109/IITCEE 57236.2023.10091034.

[8] R. B. Koyel Chakraborty and Siddhartha Bhattacharyya, "A survey of sentiment analysis on social media", IEEE Trans. Comput. Soc. Syst., vol. 4, no. 1, pp. 1-11, 2020.

[9] M. K. Sohrabi and F. Hemmatian, "An efficient preprocessing method for supervised sentiment analysis by converting sentences to numerical vectors: a twitter case study", Multimed. Tools Appl., 2019.

[10] K. M. Hasib, M. A. Habib, N. A. Towhid and M. I. H. Showrov, "A Novel Deep Learning based Sentiment Analysis of Twitter Data for US Airline Service", 2021 IEEE International Conference on Information and Communication Technology for Sustainable Development ICICT4SD 2021 - Proceedings, pp. 450-455, 2021, July.

[11] A. S. Neogi, K. A. Garg, R. K. Mishra and Y. K. Dwivedi, "Sentiment analysis and classification of Indian farmers' protest using twitter data", Int. J. Inf. Manag. Data Insights, vol. 1, no. 2, pp. 100019, 2021.

[12] B. Gaye, D. Zhang and A. Wulamu, "A tweet sentiment classification approach using a hybrid stacked ensemble technique", Inf., vol. 12, no. 9, 2021.

[13] Y. Yang, "Research and Realization of Internet Public Opinion Analysis Based on Improved TF - IDF Algorithm", 2017 16th International Symposium on Distributed Computing and Applications to Business Engineering and Science (DCABES), vol. 2018, pp. 80-83, 2017, Septe.

[14] B. Ray, A. Garain and R. Sarkar, "An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews", Appl. Soft Comput., vol. 98, no. xxxx, pp. 106935, Jan. 2021.

[15] E. K. Ampomah, Z. Qin and G. Nyame, "Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement", Information, vol. 11, no. 6, pp. 332, 2020.

[16] S. Kumar and M. Zymbler, "A machine learning approach to analyze customer satisfaction from airline tweets", J. Big Data, vol. 6, no. 1, pp. 1-16, 2019.