# Twitter Sentiment Analysis

**D. Bala Rama Krishna**, Assistant Professor

**K. Lakshmi Supraja, N. Chandana, S. Lithin Shanmukha, A.S. Ravi Teja**

*Department of CSE (AI & ML), SRK Institute of Technology, Vijayawada, A.P., India*

**ABSTRACT: -** This project addresses the problem of sentiment analysis in twitter that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral. Twitter is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters. It is a rapidly expanding service with over 200 million registered users out of which 100 million are active users and half of them log on twitter on a daily basis - generating nearly 250 million tweets per day. Due to this large amount of usage we hope to achieve a reflection of public sentiment by analyzing the sentiments expressed in the tweets. Analyzing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange.

**KEYWORDS:** - Tweets, Natural Language Processing, Random Forest, TF IDF Vectorization, Model Training.

## INTRODUCTION

This project is designed to classify tweets based on the sentiment as positive, negative and neutral. Our dataset contains four columns named as Positive, Negative, Neutral and Irrelevant. Firstly, we are going to train our model based on positive, negative and neutral comments. And later we Retrain the model by given the input by irrelevant comments. In this our Model classify the Tweets based on Double training. we used different algorithms like Random Forest, Naïve Bayas, XG Boost etc. 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 preprocess this data to eliminate any noise and subsequently use machine learning methods such as Naive Bayes, XG Boost, Random Forest, SVM. 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. Here we used the data set which contains the data of 10 different company's tweets.

## BACKGROUND

Main purpose of this project is to analyze sentiment expressed in Twitter data related to specific topics, aiming to understand public opinion, sentiment trends, and reactions to events or developments. By leveraging sentiment analysis techniques, the project seeks to provide insights for various applications such as market research, brand reputation management, and social listening. Existing methods they have used almost all the algorithms such as Random Forest, SVM, XG Boost, Vader, LSTM. In the Existing methodologies they each and every project used 3 different factors for classification, they are positive, Negative and Neutral.

Traditional methods like lexicon-based approaches and rule-based systems have limitations in scalability and context handling. Machine learning models, including Naïve Bayes and Support Vector Machines, have been widely used for sentiment classification due to their ability to learn from large datasets. Deep learning models such as RNNs, LSTMs, and CNNs have shown remarkable performance in capturing complex relationships in text data. Transfer learning has emerged as a powerful technique, leveraging pre-trained language models like BERT and GPT fine-tuned on sentiment tasks. Evaluation of sentiment analysis models relies on metrics like accuracy, precision, recall, and F1-score, with datasets like Sentiment140 and Sem Eval providing benchmarks. Challenges include sarcasm detection, context dependency, and bias mitigation, driving future research directions towards multimodal and fine-grained sentiment analysis.

## METHODOLOGY

The proposed solution introduces an innovative approach to address the challenge of Considering 4 types of factors based on the comments Positive, Negative, Neutral, Irrelevant. Our data set contains tweet comments of 10 different companies containing all the four factors, Positive, Negative, Neutral, Irrelevant. Irrelevant Data means the data or the comment is classified under any of other three such data is titled as irrelevant. Firstly, we are going to classify our models based on the three factors Positive, Negative, Neutral. Later We are giving this irrelevant data to our model to classify such tweets. So that We are giving label to that irrelevant data as Positive, Negative, Neutral. Likewise here our Models are Double trained. We are used few of the existing Models to Perform this methodology such as Random Forest, SVM, Naïve Bayes, XG Boost. We trained the models by using double training. And we tried to attain the highest accuracy among all the models we used Random Forest Got the highest accuracy 94 Percent. We focus on avoiding the Over fitting and reduce complexity etc. Random Forest: Known for its robustness and ability to handle high-dimensional data, Random Forest is an ensemble learning method that aggregates multiple decision trees to improve classification accuracy and mitigate overfitting. Naive Bayes: A simple yet effective probabilistic classifier based on Bayes' theorem, Naive Bayes assumes independence between features, making it computationally efficient and suitable for large datasets. It's particularly useful for text classification tasks like sentiment analysis. Support Vector Machine (SVM): SVM is a powerful supervised learning algorithm that constructs a hyperplane to separate classes in feature space, maximizing the margin between them. It's effective in high-dimensional spaces and is known for its versatility in handling both linear and nonlinear data. XG Boost is an optimized implementation of gradient boosting algorithms, designed for speed and performance. It sequentially builds multiple weak learners to create a strong predictive model, making it popular in various machine learning competitions and applications. Each of these models offers unique advantages in sentiment analysis tasks, from the interpretability of Naive Bayes to the robustness of Random Forest and the efficiency of SVM and XG Boost. Evaluating their performance on Twitter data can provide insights into their suitability for real-world applications.

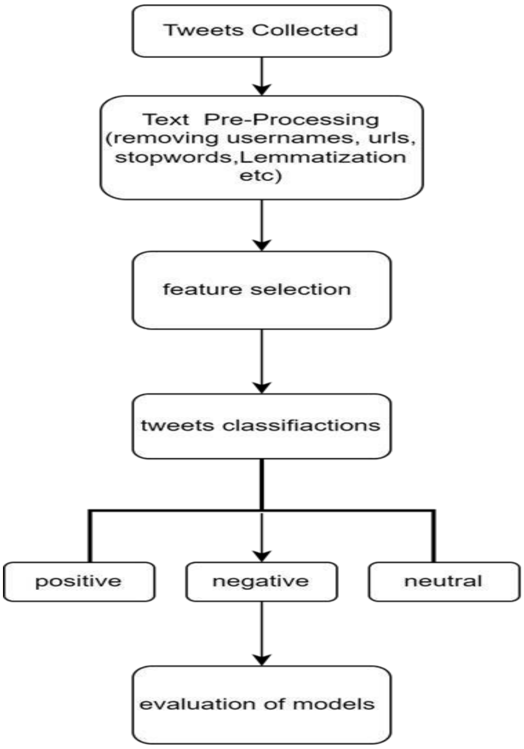
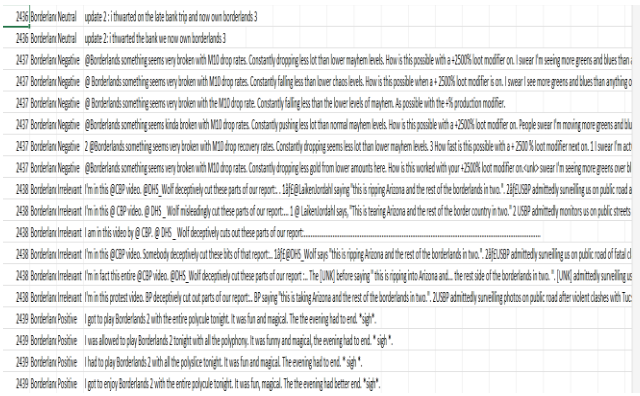


Fig 1. Methodology

**Data Collection**: This dataset was acquired from the Kagglerepository and includes tweets from 10 different Companies. In this dataset, we have 74000 rows and 4 columns.



| | | | |
|---|---|---|---|
| 2436 | Borderlanc | Neutral | update 2 : i thwarted on the late bank trip and now own borderlands 3 |
| 2436 | Borderlanc | Neutral | update 2: i thwarted the bank we now own borderlands 3 |
| 2437 | Borderlanc | Negative | @Borderlands something seems very broken with M10 drop rates. Constantly dropping less lot than lower mayhem levels. How is this possible with a +2500% loot modifier on. I swear I'm seeing more greens and blues than a |
| 2437 | Borderlanc | Negative | @ Borderlands something seems very broken with M10 drop rates. Constantly falling less than lower chaos levels. How is this possible when a + 2500% loot modifier is on. I swear I see more greens and blues than anything o |
| 2437 | Borderlanc | Negative | @ Borderlands something seems very broken with the M10 drop rate. Constantly falling less than the lower levels of mayhem. As possible with the +% production modifier. |
| 2437 | Borderlanc | Negative | @Borderlands something seems kinda broken with M10 drop rates. Constantly pushing less lot than normal mayhem levels. How is this possible with a +2500% loot modifier on. People swear I'm moving more greens and blu |
| 2437 | Borderlanc | Negative | 2 @Borderlands something seems very broken with M10 drop recovery rates. Constantly dropping seems less lot than lower mayhem levels. 3 How fast is this possible with a + 2500 % loot modifier next on. 1 I swear I'm act |
| 2437 | Borderlanc | Negative | @Borderlands something seems very broken with M10 drop rates. Constantly dropping less gold from lower amounts here. How is this worked with your +2500% loot modifier on.<unk> swear I'm seeing more greens over bl |
| 2438 | Borderlanc | Irrelevant | I'm in this @CBP video. @DHS_Wolf deceptively cut these parts of our report:. . 1âƒ£@LaikenJordahl saying "this is ripping Arizona and the rest of the borderlands in two.". 2âƒ£USBP admittedly surveilling us on public road a |
| 2438 | Borderlanc | Irrelevant | I'm in this @ CBP video. @ DHS _ Wolf misleadingly cut these parts of our report:... 1 @ LaikenJordahl says, "This is tearing Arizona and the rest of the border country in two." 2 USBP admittedly monitors us on public streets |
| 2438 | Borderlanc | Irrelevant | I am in this video by @ CBP. @ DHS _ Wolf deceptively cuts out these parts of our report:.................................................................................................... |
| 2438 | Borderlanc | Irrelevant | I'm in this @CBP video. Somebody deceptively cut these bits of that report:.. 1âƒ£@DHS_Wolf says "this is ripping Arizona and the rest of the borderlands in two.". 2âƒ£USBP admittedly surveilling us on public road of fatal cl |
| 2438 | Borderlanc | Irrelevant | I'm in fact this entire @CBP video. @DHS_Wolf deceptively cut these parts of our report :.. The [UNK] before saying " this is ripping into Arizona and... the rest side of the borderlands in two. ". [UNK] admittedly surveilling us |
| 2438 | Borderlanc | Irrelevant | I'm in this protest video. BP deceptively cut out parts of our report:.. BP saying "this is taking Arizona and the rest of the borderlands in two.". 2USBP admittedly surveilling photos on public road after violent clashes with Tuc |
| 2439 | Borderlanc | Positive | I got to play Borderlands 2 with the entire polycule tonight. It was fun and magical. The the evening had to end. *sigh*. |
| 2439 | Borderlanc | Positive | I was allowed to play Borderlands 2 tonight with all the polyphony. It was funny and magical, the evening had to end. * sigh *. |
| 2439 | Borderlanc | Positive | I had to play Borderlands 2 with all the polyslice tonight. It was fun and magical. The evening had to end. * sigh *. |
| 2439 | Borderlanc | Positive | I got to enjoy Borderlands 2 with the entire polycule tonight. It was fun, magical. The the evening had better end. *sigh*. |

Fig 2: Data Set

**Preprocessing**: For the model to be more accurate, we must preprocess the data. To preprocess the tweets collected, wefirst remove the usernames, URLs, stop words, and so on. After we have removed all of the usernames, URLs, and stop words, we tokenize the text and utilize the lemmatize approach to reduce the term to its root form.

**Machine Learning Models**:

**1.** Random Forest: A machine learning algorithm that integrates the opinions of several "trees" (individual models) to improve predictions, resulting in a stronger and better-performing model. It is user-friendly and versatile, capable of efficiently handling both classification and regression issues. Random Forest algorithm is a powerful tree learning technique in Machine Learning**.** It works by creating a number of Decision Trees during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition.

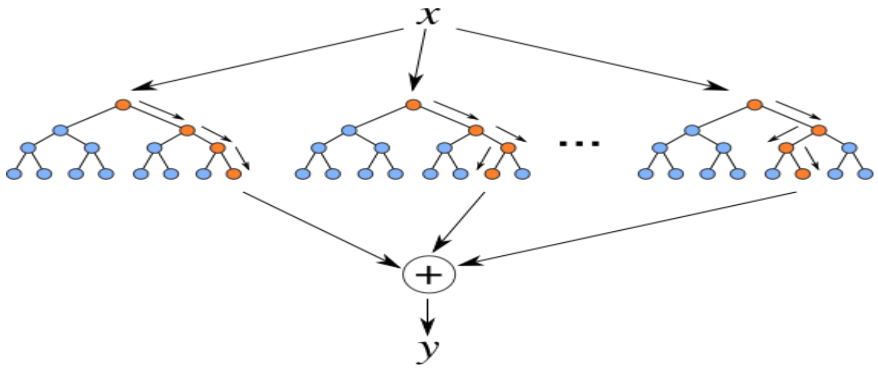


Fig 3. Random Forest

**2.** XG Boost: An ensemble learning method used for supervised learning problems like regression and classification. It creates a predictive model by iteratively merging the predictions of numerous independent models, most often decision trees. XG Boost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XG Boost stands for "Extreme Gradient Boosting" and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression.

**3.** Naïve Bayes: A Naive Bayes classifier, a family of algorithms based on Bayes' Theorem. Despite the "naive" assumption of feature independence, these classifiers are widely utilized for their simplicity and efficiency in machine learning. The article delves into theory, implementation, and applications, shedding light on their practical utility despite oversimplified assumptions. Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

SVM: Support Vector Machine (SVM) is a powerful machine learning algorithm used for linear or nonlinear classification, regression, and even outlier detection tasks. SVMs can be used for a variety of tasks, such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection. SVMs are adaptable and efficient in a variety of applications because they can manage high-dimensional data and nonlinear relationships.

## IV Results and Discussions

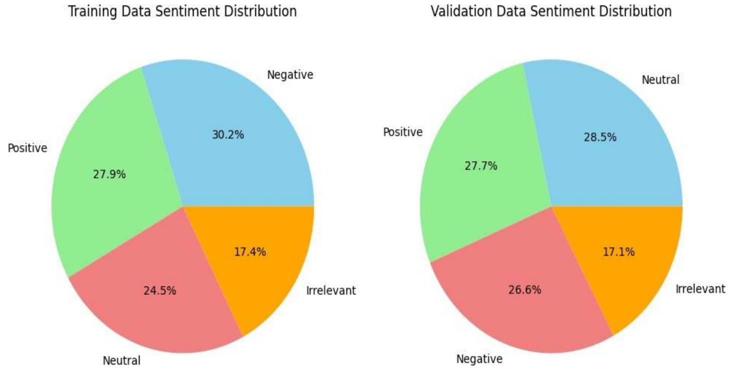


Fig 4: Distribution of sentiment

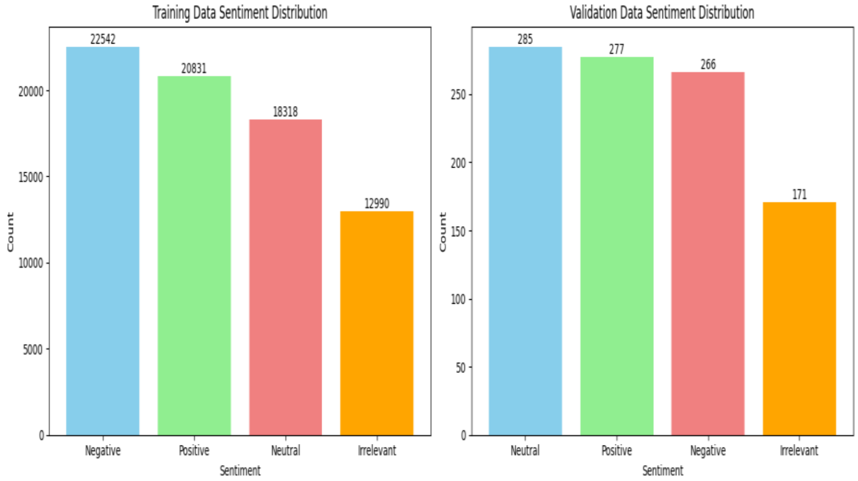


Fig 5: Represent of Bar Graph

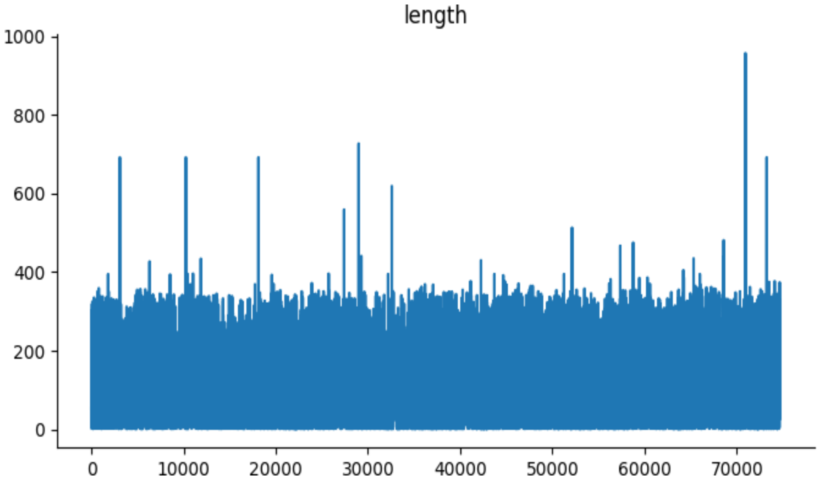


Fig:6 Distribution of Length of Comment

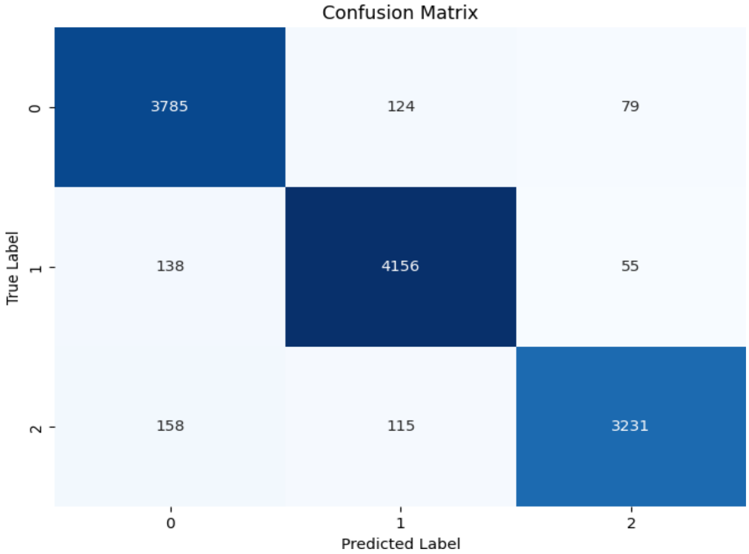


Fig 7: Confusion Matrix of Random Forest

Fig 8: Word Cloud

## IV SYSTEM ARCHITECTURE

The system architecture consists of several components, including the user interface, database storage, and c. Firstly, there's the part you see and interact with - the User Interface. It's like the front door to the whole system. You use it to Select the Model in which you want to classify the Tweet or Comment. You can select the Model like for example Random Forest. Later Write the comment as your wish Our interface will display the output whether it is positive, Negative, Neutral.

The system architecture comprises data collection from Twitter's API, followed by preprocessing steps like text normalization and feature extraction. Processed data is then fed into multiple sentiment analysis models including Random Forest, Naive Bayes, SVM, and XG Boost for classification. Each model generates sentiment predictions which are aggregated or compared to determine the final sentiment label. The architecture allows for flexibility in incorporating additional models or refining existing ones based on performance metrics. Deployment of the system can be achieved through a web interface or API endpoints for real-time or batch sentiment analysis. Scalability considerations ensure the system can handle large volumes of Twitter data efficiently. Modular design facilitates easy integration with other components or systems for broader analytics purposes.

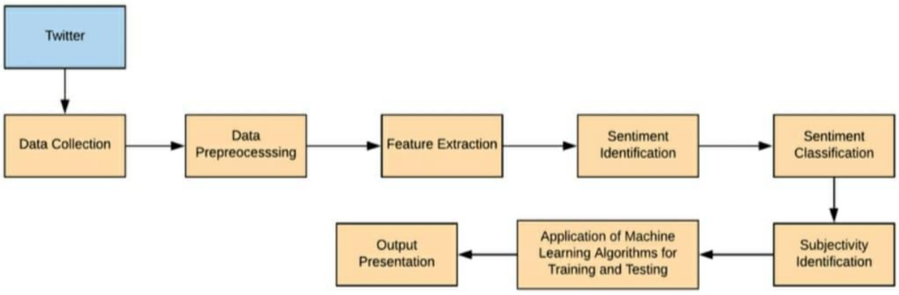


Fig-9: Architecture Diagram

## V APPLICATIONS

1. **Real-time monitoring:** It allows stakeholders to address negative opinions quickly whilecapitalizing on good sentiment.
2. **Audience Engagement:** Marketing teams may adapt their efforts based on sentiment analysisdata to better connect with their target audience.
3. **Brand impression:** Sponsors can assess their brand's impression among fans and alter strategy appropriately.
4. **Fan Experience Enhancement:** Event organizers may use sentiment analysis to identify areas for improvement and then improve the entire fan experience, resulting in maximum satisfaction.

**5.Crisis Management:** Quickly identify possible controversies or unfavorable situations and take corrective steps to limit their impact.

## VI CONCLUSION

This Project Twitter sentiment analysis effectively analyzed a large volume of tweets to gauge public sentiment on a particular topic or event. Through natural language processing techniques, we categorized tweets as positive, negative, or neutral, providing valuable insights into public opinion. By employing machine learning algorithms, we achieved high accuracy in sentiment classification, enabling us to identify trends, sentiments, and key influencers within the Twitter sphere. This project demonstrates the power of data analytics in understanding public sentiment and its potential applications in various domains such as marketing, politics, and public opinion research." The utilization of various machine learning methodologies, including XG Boost, Random Forest, Naïve bayes, SVM for performing sentiment analysis on Twitter. Naïve Bayes Accuracy is 80.5%, XG boost Accuracy is 74.44%, Support Vector Machine Accuracy is 74.43% Random Forest |Accuracy is 94.20%, Among all of these Random Forest have achieved the highest accuracy of 94.20 percent. We also double trained the model in this project. Additionally, sentiment scores will be calculated for each tweet by leveraging these models, encompassing dimensions of positivity, negativity, and neutrality. Lastly, a comparative study among these models will be conducted. In the future, the Neural Network model shows promise and has the potential to outperform the other models in terms of accuracy if it is fine-tuned.

## VII Future Scope

In the Present Existing models are trained with the Machine learning Models. But there is scope that in future the models can be trained with the Deep Learning techniques. May be by using Neural Networks. Also, with advancement in natural language processing and mission learning, we could explore real time sentiment analysis, multilingual sentiment analysis, also sentiment analysis for other Social Media Platform. Explore more sophisticated Neural Network Architecture, Recurrent Neural Network, LSTM Etc.

## VIII References

[1] Dhanta, R., Sharma, H., Kumar, V. & Singh, H. O. (2023). Twitter sentimental analysis using machine learning. International Journal of Communication and Information Technology, 4(1), 71–83. DOI: 10.33545/2707661x.2023.v4.i1a.63.

[2] Bahrawi, N. (2019). Sentiment analysis using random forest algorithm-online social media based. Journal of Information Technology and Its Utilization,2(2),29. https://doi.org/10.30818/jitu.2.2.2695.

[3] Jacob, S. S. & Vijayakumar, R. (2021). Sentimental analysis over twitter data using clustering-based machine learning algorithm. Journal of Ambient Intelligence and Humanized Computing. https://doi.org/10.1007/s12652-020- 02771-9.

[4] Ravi Kumar, G., Venkata Sheshanna, K. & Anjan Babu, G. (2021). Sentiment analysis for airline tweets utilizing machine learning techniques. In: EAI/Springer Innovations in Communication and Computing, pp. 791–799. Springer Science and Business Media Deutschland GmbH.

https://doi.org/10.1007/978-3-030-49795-8_75.

[5] Khan L, Amjad A, Afaq KM & Chang H-T. (2022). Deep sentiment analysis using CNNLSTM architecture of English and roman Urdu text shared in social media. Applied Sciences, 12(5), 2694. https://doi.org/10.3390/app12052694.

[6] Alexander Pak & Patrick Paroubek. (2010). Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association.

[7] Kumar, D. & Rao, S. (2020). A sentiment analysis of twitter data using bi-directional long short-term memory. DOI: 10.1007/978-3-030-30271- 9_16.

[8] Shobana, G., Vigneshwara, B. & Maniraj Sai, A. (2019). Twitter sentimental analysis. International Journal of Recent Technology and Engineering,7(4),343–346. https://doi.org/10.46501/ijmtst061266

[9] Dashrath Mahto, Subhash Chandra Yadav & Gotam Singh Lalotra. (2022). Sentiment prediction of textual data using hybrid convbidirectional-lstm model. Mobile Information Systems. https://doi.org/10.1155/2022/1068554.

[10 I. Guellil & K. Boukhalfa. (2015). Social big data mining: A survey focused on opinion mining and sentiments analysis. 12th International Symposium on Programming and Systems (ISPS), Algiers.

[11] A. Sweelinck, D. Haczyk & M. Haczyk. (2023). Graph neural networks for natural language processing in human-robot interaction. Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, Poland, pp.

[12] K. S. Madhu, B. C. Reddy, C. Damarukanadhan, M. Polireddy & N. Ravinder. (2021). Real time sentimental analysis on twitter. 6th International Conference on Inventive Computation Technologies, Coimbatore, India, pp. 1030-1034. DOI:10.1109/ICICT50816.2021.9358772.

[13] N. Deepa, J. S. Priya & T. Devi. (2023). Sentimental analysis recognition in customer review using Novel-CNN. International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, pp. 1-4. doi: 10.1109/ICCCI56745.2023.10128627.

[14] Çilgin, C., Baş, M., Bilgehan, H. & Ünal, C. (2022). Twitter sentiment analysis during covid19 outbreak with VADER. AJIT-e: Academic Journal of Information Technology, 13(49), 72– 89. https://doi.org/10.5824/ajite.2022.02.001.x

[15] N. Deepa, J. S. Priya & T. Devi. (2023). Sentimental analysis recognition in customer review using Novel-CNN. International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, pp. 1-4. doi: 10.1109/ICCCI56745.2023.10128627

[16] Y. E. Cakra & B. Distiawan Trisedya. (2015). Stock price prediction using linear regression based on sentiment analysis. International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, pp. 147-154. DOI: 10.1109/ICACSIS.2015.7415179.

[17] P. Khurana Batra, A. Saxena, Shruti & C. Goel. (2020). Election result prediction using twitter sentiments analysis. Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, pp. 182- 185. DOI: 10.1109/PDGC50313.2020.9315789

[18] A. Z. Adamov & E. Adali. (2016). Opinion mining and Sentiment Analysis for contextual online-advertisement. IEEE 10th International Conference on Application of Information and Communication Technologies (AICT), Baku, Azerbaijan, pp. 1-3. DOI: 10.1109/ICAICT.2016.7991682.