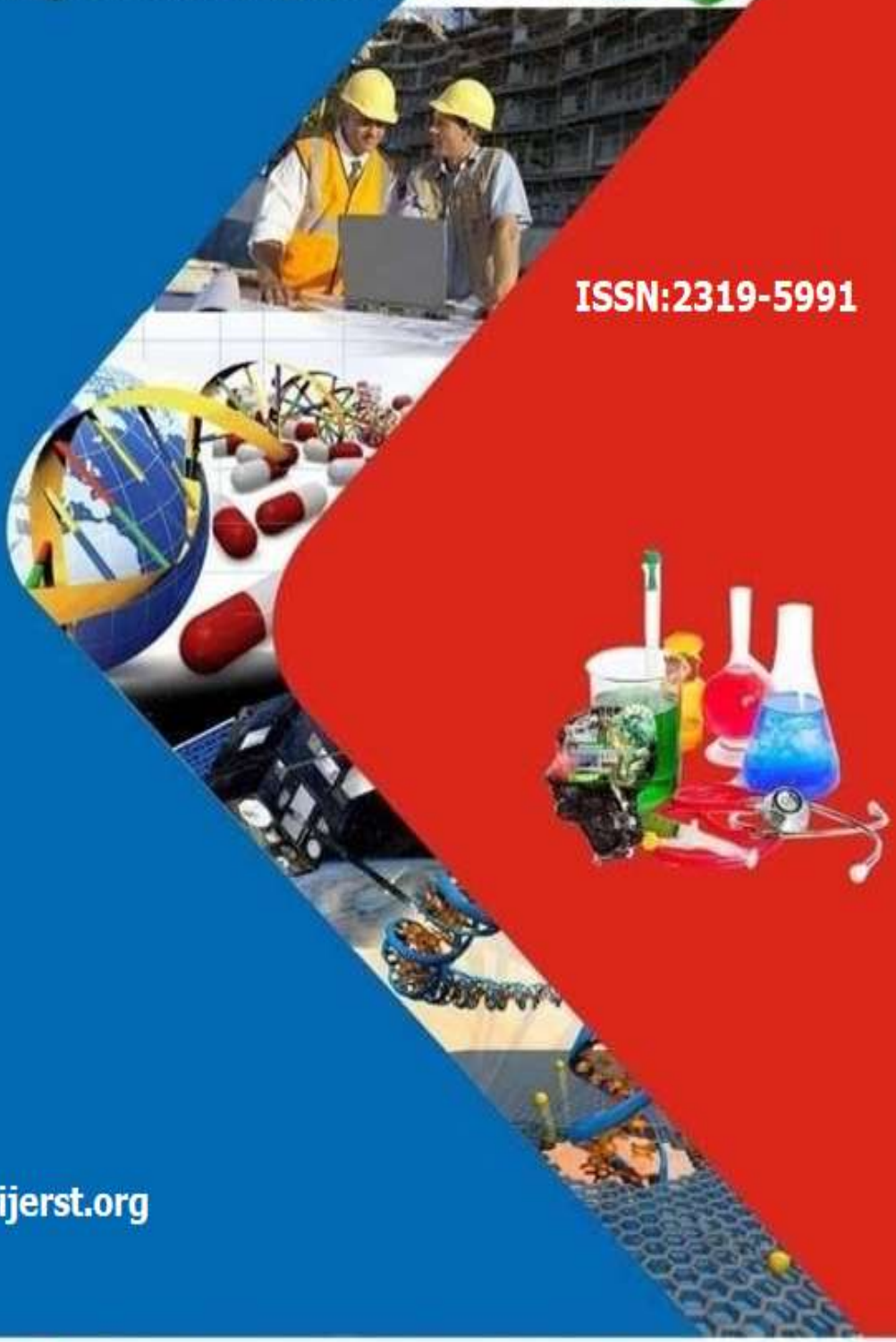
International Journal of Engineering Research and Science & Technology

IJERST

ISSN:2319-5991



www.ijerst.org

E-mail: editor@ijerst.org or ijerst.editor@gmail.com

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

UNVEILING INSIGHTS WITH TWITTER DATA: EXPLORING TRENDS, SENTIMENTS, AND PREDICTIONS THROUGH SOCIAL MEDIA MINING

S. Sundeep kumar[1*], Himaja[2], Jella Siddartha[2], M. Venkata Sree Nayan[2], L. Bhadru[2]

[1] Assistant Professor, [2]UG Student, [1,2] Department of Computer Science and Engineering (AIML)

[1,2]Sree Dattha Institute of Engineering and Science, Sheriguda, Hyderabad, Telangana

ABSTRACT

With the rapid rise of social media, Twitter has emerged as a valuable source of real-time data, offering insights into public opinion, sentiments, and trending topics. The unstructured and noisy nature of Twitter content—characterized by hashtags, mentions, abbreviations, and emoticons—poses significant challenges for traditional text processing techniques like tokenization and stemming, which often fall short in capturing the platform's linguistic nuances. As the relevance of Twitter data grows in areas such as sentiment analysis, brand monitoring, and trend prediction, the need for an advanced and comprehensive pre-processing approach becomes crucial. Effective preprocessing helps filter out irrelevant information while preserving context, enabling machine learning algorithms to classify tweets more accurately. This research focuses on leveraging machine learning for Twitter data classification by introducing a robust pre-processing pipeline, ultimately contributing to improved accuracy, deeper understanding of public sentiment, and more informed decision-making for businesses and researchers alike.

Keywords: Twitter Data, Sentiment Analysis, Preprocessing, Machine Learning, Public Opinion

1.INTRODUCTION

The history of analyzing Twitter data for insights traces back to the early 2000s when social media platforms began to burgeon. With the inception of Twitter in 2006, a new avenue for real-time data analysis emerged. Initially, researchers and businesses viewed Twitter as a platform for social interaction[1,2,3]. However, as its user base expanded exponentially, it became evident that Twitter harbored a wealth of information beyond mere conversations. Around 2010, the academic community and industry pioneers started recognizing Twitter's potential as a goldmine for understanding public sentiment, predicting trends, and conducting market research[4,5,6]. This recognition marked the onset of a concerted effort to develop methods and algorithms specifically tailored for processing and classifying Twitter data effectively. Traditional systems relied on rudimentary text processing techniques like tokenization, stemming, and stop-word removal. These methods, while useful, struggled to cope with the unique characteristics of Twitter data, such as hashtags, mentions, and emoticons. Consequently, researchers began to explore more sophisticated approaches[7] to address the challenges posed by the unstructured and noisy nature of Twitter data. The evolution of machine learning algorithms further propelled the analysis of Twitter data. Researchers started experimenting with various models to extract insights from the vast pool of tweets generated every second. This experimentation led to the development of novel techniques aimed at improving the accuracy and efficiency of Twitter data classification.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

2. LITERATURE SURVEY

Sanjay et al. [8] conducted sentiment analysis on Twitter data related to the Indian farmer protests to gain insights into global public sentiment. They employed algorithms to analyze approximately twenty thousand tweets associated with theprotests and assess the sentiments expressed. The researchers analyzed and contrasted the success of 2 popular text representation techniques BoW and TF-IDF, and discovered that BoW outperformed TF-IDF in sentiment analysis accuracy. The study further involved the application of various classifiers, including SVM, RF, DT, and NB, on the dataset. The results revealed that the RF classifier achieved the best possible accuracy among the evaluated classifiers.

Behl et al. [9] gathered tweets related to various natural disasters and categorized them into three groups based on their content: "resource availability," "resource requirements," and "others." To accomplish this classification task, they employed a Multi-Layer Perceptron (MLP) network with an optimizer. The proposed model demonstrated an accuracy of 83%, indicating its effectiveness in accurately classifying the tweets into the designated categories.

Tan et al. [10] introduced a model that combined BI-LSTM, RoBERTa, and GRU models. To further enhance the general effectiveness of sentiment analysis, the model's predictions were averaged using majority voting. Addressing the challenges posed by unbalanced datasets, the researchers enhanced the data by utilizing GloVe pre-trained word embeddings. The experimental results demonstrated that the proposed model surpassed state-of-the-art approaches, achieving accuracy rates of 0.942, 0.892, and0.9177 on the Sentiment 140, USAirlines, and IMDB datasets, respectively.For Aspect-level SA, Lu et al. [11] presented IRAN (Interactive Rule Attention Network). Tosimulate the operation of grammar at the sentence level, IRAN includes a grammar rule encoder that normalizes the result of adjacent locations. Furthermore, IRAN makes use of an attention network that interacts with its environment to better understand the target and its surroundings. We show that IRAN learns informative features successfully and beats baseline models by experimenting on the ACL 2014 Twitter & SemEval 2014 datasets. As a result of these results, it is clear that IRAN is an effective tool for aspect-level sentiment analysis, which can lead to enhanced performance in the field.

In their study, Mehta et al. [12] conducted a relative investigation of SA specifically focused on big data. They identified six types of emotions, namely happy, sad, joy, surprise, disgust, and fear. Additionally, they judged various methods for emotion identification that can serve as potential avenues for future research in the field. This analysis provides valuable insights into sentiment analysis in the context of big data, offering a foundation for exploring emotion identification techniques and theirapplications.

He et al. [13] introduced LGCF, a multilingual learning paradigm that emphasized active learning inboth global and local contexts. Unlike its predecessors, this model, LGCF International Journal of Intelligent Systems and Applications in EngineeringIJISAE, 2024, 12(1), 235–266|237demonstrated the ability to effectively learn the connections between target aspects and local contexts, along with the connections between target aspects and global contexts simultaneously. This innovative approach enables the model to capture and utilizeboth local and global contextual information efficiently, enhancing its overall performance in sentiment analysis tasks.

In their study [14], an extensive evaluation of sentiment polarity classification methods was specifically designed for Twitter text. Notably, they expanded the comparison by including a

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

combination of classifiers in their analysis and introduced the aggregation and utilization of manually annotated tweets for method evaluation. This aspect is considered a significant contribution because previous attempts at automated annotation based on features like emoticons have proven problematic. Such automated approaches often fail to accurately capture the overall sentiment expressed by the author, particularly when considering instances of neutral sentiment or the absenceof sentiment in the text. The inclusion of manual annotations addresses this limitation and adds value to the evaluation process of SA methods for Twitter text.

To better understand the state of the art in SA using DNNs and CNNs, Qurat et al. [15] undertook a systematic literature review of current studies. Topics covered in their investigation of sentiment analysis included text sentiment categorization, cross-lingual analysis, and both textual and visual analysis. Datasets were culled from a wide range of social media platforms. The authors presented the various stages of the successful construction of DL models in emotion analysis and noted that many difficulties in this field were efficiently solved with high accuracy using deep learning methodologies. With their more complex structures, deep learning networks were able to extract and represent features more accurately than traditional neural networks and SVMs. This study demonstrates the benefits of using DL models for sentiment analysis, which can lead to improved results in emotion analysis.

3.PROPOSED SYSTEM

- Importing Libraries:
  The script begins by importing necessary libraries, including NumPy, Pandas, Matplotlib, Seaborn, NLTK, and warnings.
- Loading Data:
  The training and testing datasets are loaded from CSV files using Pandas.
  The shape of the datasets is printed to provide an overview of the data size.
- Exploratory Data Analysis (EDA):
  Displaying the first few rows of the training and testing datasets to inspect the structure of the data. Checking for missing values in both datasets. Exploring positive and negative comments in the training set. Visualizing the distribution of tweet lengths in both training and testing datasets. Creating a new column to represent the length of each tweet.Grouping the data by label (positive or negative) and analyzing statistics.
- Data Visualization:
- Creating count plots and histograms to visualize the distribution of tweet lengths, label frequencies, and hashtag frequencies. Generating word clouds to display the most frequent words in the overall vocabulary, neutral words, and negative words.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

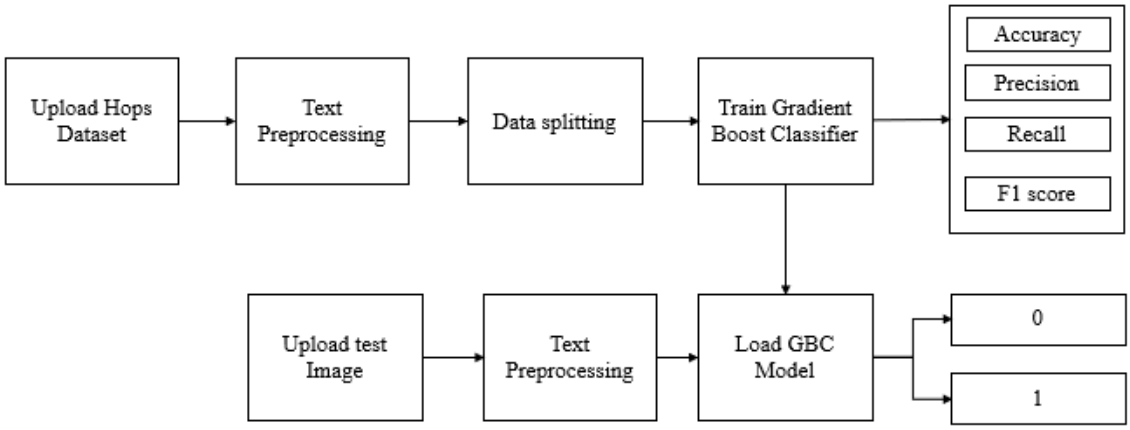


Figure 1: Block Diagram

- Hashtag Analysis:
  Extracting hashtags from both positive and negative tweets. Creating frequency distributions and bar plots to display the most common hashtags in each category.
- Word Embeddings with Word2Vec:
  Using Gensim to train a Word2Vec model on tokenized tweets. Demonstrating word similarities for certain words using the trained Word2Vec model.
- Text Preprocessing:
  Removing unwanted patterns, converting text to lowercase, and stemming words using NLTK.Creating bag-of-words representations for both the training and testing datasets.
- Model Training:
  Splitting the training dataset into training and validation sets.
- Standardizing the data using StandardScaler.
  Training machine learning models including RandomForestClassifier, LogisticRegression
- Evaluating the models on the validation set, calculating training and validation accuracy, F1 score, and generating confusion matrices. The script covers a wide range of tasks from data loading and exploration to text preprocessing, visualization, and training various machine learning models for sentiment analysis on Twitter data.

3.1 Gradient Boosting Classifier

Gradient Boosting Classifier (GBC) is a powerful machine learning algorithm that builds an ensemble of weak learners, usually decision trees, and combines them sequentially to minimize a loss function. It is a boosting technique where each new tree corrects the errors made by the previous ones.

Architecture:

1. **Loss Function**: The model optimizes a loss function, which could be binary cross-entropy (for classification) or mean squared error (for regression).
2. **Weak Learners**: It typically uses shallow decision trees (stumps) as weak learners. Each tree corrects the mistakes of the previous ones by focusing on instances with higher residual errors.
3. **Gradient Updates**: After each iteration, the gradient of the loss function is calculated to update the model parameters.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

4. **Shrinkage**: To prevent overfitting, a learning rate is applied to the weights of the trees to control the contribution of each tree.

Advantages:

- **High Accuracy**: GBC often outperforms other ensemble models in terms of accuracy, especially on structured/tabular data.
- **Robustness to Overfitting**: With appropriate regularization techniques like shrinkage (learning rate) and early stopping, GBC is less prone to overfitting compared to standard decision trees.
- **Handles Imbalanced Data Well**: GBC can be tailored to handle imbalanced datasets by adjusting the loss function or class weights.

4. RESULTS AND DISCUSSION

The provided dataset appears to be related to Twitter data, containing information such as tweet IDs, labels, and the content of the tweets. Here's a detailed description:

- ID Column: Represents the unique identifier for each tweet.
- Label Column: Indicates whether the tweet is labeled as 0 or 1. In this context, it seems to be a binary classification, with 0 possibly representing non-dysfunctional content and 1 representing dysfunctional content.
- Tweet Column: Contains the actual text content of the tweets. The tweets seem to vary in topics, including mentions of Lyft, birthday wishes, expressions of love, motivational content, and more.

It's important to note that without additional context, the specific criteria for labeling tweets as dysfunctional or the context behind the labeling are not clear. The dataset comprises a diverse range of tweets, suggesting it may be used for sentiment analysis, classification, or related natural language processing tasks.

Figure 1: Data frame used for Twitter data analysis figure likely represents the structure and content of the data frame used for Twitter data analysis. It might include information such as tweet text, sentiment labels, other relevant features.

Figure 2: Count plot of target column figure is a visual representation, likely in the form of a bar chart, showing the distribution or count of different classes in the target column. In the context of Twitter data analysis, the target column might represent sentiment labels such as positive, negative, or neutral.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation
...	...	...	...
31957	31958	0	ate @user isz that youuu?ð□□□ð□□□ð□□□ð□□□ð□□□ð...
31958	31959	0	to see nina turner on the airwaves trying to...
31959	31960	0	listening to sad songs on a monday morning otw...
31960	31961	1	@user #sikh #temple vandalised in in #calgary,...
31961	31962	0	thank you @user for you follow

31962 rows × 3 columns

Figure 2: Data frame used for Twitter data analysis

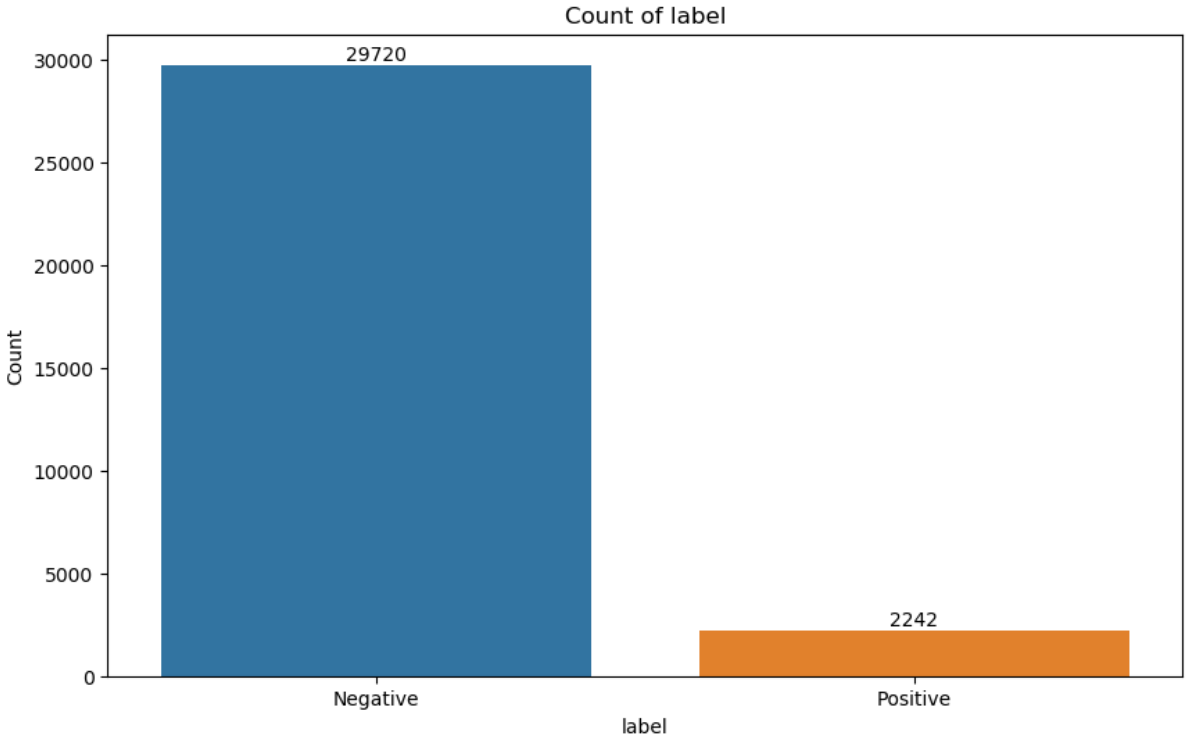


Figure 3: Count plot of target column

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

```
Model loaded successfully.
ExtraTreesClassifier Accuracy    : 92.86682657211388
ExtraTreesClassifier Precision   : 50.0
ExtraTreesClassifier Recall      : 46.43341328605694
ExtraTreesClassifier FSCORE      : 48.15075159511192

 ExtraTreesClassifier classification report
              precision    recall  f1-score   support

    Negative       0.93      1.00      0.96      8905
    Positive       0.00      0.00      0.00       684

    accuracy                           0.93      9589
   macro avg       0.46      0.50      0.48      9589
weighted avg       0.86      0.93      0.89      9589
```

Figure 4: ETC Classification report

Figure 4 shows

- **Accuracy:** 92.87% This indicates that the model correctly predicted 92.87% of the data points in the test set. It's a general measure of overall performance.
- **Precision:** 50.0% This metric measures the proportion of positive predictions that were actually correct. In other words, out of all the instances the model predicted as positive, 50% were truly positive.
- **Recall:** 46.43% This metric measures the proportion of actual positive instances that were correctly predicted. It indicates how well the model was able to identify all the positive cases.
- **F1-score:** 48.15% This metric combines precision and recall into a single value. It provides a balanced measure of both metrics, considering both the model's ability to correctly predict positive instances and its ability to avoid false positives.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

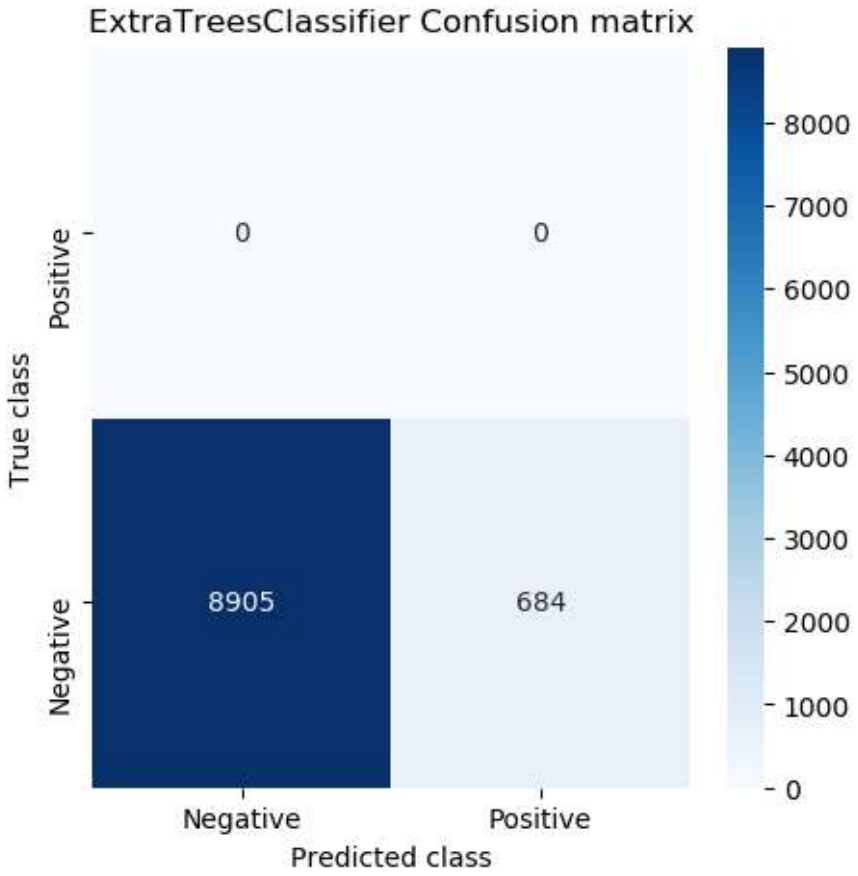


Figure 5: ETC Confusion Matrix

The confusion matrix Fig 5 for the ExtraTreesClassifier model shows the distribution of predicted and actual classes. The diagonal elements represent correct predictions (e.g., 8905 Negative instances were correctly predicted as Negative). Off-diagonal elements indicate incorrect predictions (e.g., 684 Positive instances were incorrectly predicted as Negative). The color intensity of each cell corresponds to the number of instances in that category, with darker colors indicating larger quantities.

```
Model loaded successfully.
GradientBoostingClassifier Accuracy    : 94.69183439357597
GradientBoostingClassifier Precision   : 65.22184297539657
GradientBoostingClassifier Recall      : 90.18097278669228
GradientBoostingClassifier FSCORE      : 71.26905254116195

 GradientBoostingClassifier classification report
              precision    recall  f1-score   support

    Negative       0.95      1.00      0.97      8905
    Positive       0.85      0.31      0.45       684

    accuracy                           0.95      9589
   macro avg       0.90      0.65      0.71      9589
weighted avg       0.94      0.95      0.94      9589
```

Figure 6: GBC Classification report

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

Figure 6 shows that

- **Accuracy:** This is the overall correctness of the model. It's calculated as the number of correct predictions divided by the total number of predictions. In this case, the accuracy is 94.69%, which means the model correctly predicted 94.69% of the samples.
- **Precision:** This measures how many of the positive predictions made by the model were actually correct. It's calculated as the number of true positives divided by the sum of true positives and false positives. In this case, the precision is 65.22%, which means that out of all the samples the model predicted as positive, only 65.22% were truly positive.
- **Recall:** This measures how many of the actual positive samples the model correctly identified. It's calculated as the number of true positives divided by the sum of true positives and false negatives. In this case, the recall is 90.18%, which means that the model correctly identified 90.18% of the positive samples.
- **F1-score:** This is a harmonic mean of precision and recall. It provides a balance between precision and recall. A higher F1-score indicates better overall performance. In this case, the F1-score is 71.27%, which is a good balance between precision and recall.

Classification report: This table provides a more detailed breakdown of the model's performance for each class (negative and positive). It includes precision, recall, F1-score, and support for each class. The model achieved high accuracy (94.69%) but had some limitations in precision (65.22%) and recall (90.18%). The F1-score of 71.27% indicates a reasonable balance between precision and recall. The classification report provides further insights into the model's performance for each class.

5.CONCLUSION

With the advancement of web technology and its growth, there is a huge volume of data present on the web for internet users and a lot of data is generated too. The Internet has become a platform for online learning, exchanging ideas and sharing opinions. Social networking sites like Twitter, Facebook, Google+ are rapidly gaining popularity as they allow people to share and express their views about topics, have discussions with different communities, or post messages across the world. Therefore, this project implemented the sentiment analysis of twitter dataset for opinion mining using NLP, AI, and lexicon-based approaches, together with evaluation metrics. Using various machine learning algorithms like Naive Bayes, and logistic regression, this work provided research on twitter data streams. In addition, this project has also discussed general challenges and applications of Sentiment Analysis on Twitter.

REFERENCES

[1] Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. International Journal of Information Management Data Insights, 1(2), 100019. https://doi.org/10.1016/j.jjimei.2021.100019.

[2] Behl, S., Rao, A., Aggarwal, S., Chadha, S., & Pannu, H. (2021). Twitter for disaster relief through sentiment analysis for COVID-19 and natural hazard crises. International Journal of Disaster Risk Reduction, 55, 102101. https://doi.org/10.1016/j.ijdrr.2021.102101.

[3] Tan, K. L., Lee, C. P., Lim, K. M., & Anbananthen, K. S. M. (2022). Sentiment Analysis With Ensemble Hybrid Deep Learning Model. IEEE Access, 10, 103694–103704. https://doi.org/10.1109/access.2022.3210182.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

[4] Lu, Q., Zhu, Z., Zhang, D., Wu, W., & Guo, Q. (2020). Interactive Rule Attention Network for Aspect-Level Sentiment Analysis. IEEE Access, 8, 52505-52516,, https://doi.org/10.1109/ACCESS.2020.2981139.

[5] Mehta, K & Panda, S. (2019). A Comparative Analysis Of Sentiment analysis In Big Data. International Journal of Computer Science and Information Security, 17, 31-40.

[6] J He, J., Wumaier, A., Kadeer, Z., Sun, W., Xin, X., & Zheng, L. (2022). A Local and Global Context Focus Multilingual Learning Model for Aspect-Based Sentiment Analysis. IEEE Access, 10, 84135–84146. https://doi.org/10.1109/access.2022.3197218.

[7] E. Psomakelis, K. Tserpes, D. Anagnostopoulos, and T. Varvarigou, "Comparing methods for twitter sentiment analysis," KDIR 2014 -Proceedings of the Int. Conf. on Knowledge Discovery and Information Retrieval, pp. 225-232, 2014.

[8] Qurat Tul Ain_, Mubashir Ali_, Amna Riazy, Amna Noureenz, Muhammad Kamranz, Babar Hayat_ and A. Rehman, Sentiment Analysis Using Deep Learning Techniques: A Review , International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, 2017.

[9] A. Lopez-Chau, D. Valle-Cruz, and R. Sandoval-Almaz ´ an, "Sentiment ´ Analysis of Twitter Data Through Machine Learning Techniques," Software Engineering in the Era of Cloud Computing, pp. 185–209, 2020. Publisher: Springer, Cham.

[10] P. Kalaivani and D. Dinesh, "Machine Learning Approach to Analyze Classification Result for Twitter Sentiment," in 2020 International Conference on Smart Electronics and Communication (ICOSEC), (Trichy, India), pp. 107–112, IEEE, Sept. 2020.

[11] A. B. S, R. D. B, R. K. M, and N. M, "Real Time Twitter Sentiment Analysis using Natural Language Processing," International Journal of Engineering Research & Technology, vol. 9, July 2020. Publisher: IJERT-International Journal of Engineering Research & Technology.J. Ranganathan and A. Tzacheva, "Emotion Mining in Social Media Data," Procedia Computer Science, vol. 159, pp. 58–66, Jan. 2019.S. Xiong, H. Lv, W. Zhao, and D. Ji, "Towards Twitter sentiment classification by multi-level sentiment-enriched word embeddings," Neurocomputing, vol. 275, pp. 2459–2466, Jan. 2018.

[12] S. Aloufi and A. E. Saddik, "Sentiment Identification in Football-Specific Tweets," in IEEE Access, vol. 6, pp. 78609-78621, 2018, doi: 10.1109/ACCESS.2018.2885117.

[13] S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri, "Sentiment Analysis of Twitter Data," 2019 International Conference on Computer and Information Sciences (ICCIS), 2019, pp. 1-4, doi: 10.1109/ICCISci.2019.8716464.

[14] Arora, M., Kansal, V. Character level embedding with deep convolutional neural network for text normalization of unstructured data for Twitter sentiment analysis. Soc. Netw. Anal. Min. 9, 12 (2019). https://doi.org/10.1007/s13278-019-0557-y

[15] L. Wang, J. Niu and S. Yu, "SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 10, pp. 2026-2039, 1 Oct. 2020, doi: 10.1109/TKDE.2019.2913641.