iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70023

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Twitter Insight: A Comprehensive Pre-processing Approach for Twitter Sentiment Analysis

P. Yashwanth[1], P. Shashank[2], M. Prakash[3], P. Mallikarjun[4]
Sreenidhi Institute of Science and Technology, Hyderabad 501301, India

Abstract: ***The vast expansion of online news information in contemporary stimes requires efficient systems for classifying content while determining underlying emotional sentiments. The proposed integrated system uses Natural Language Processing methods to both sort news articles into designated categories political, sports, business and entertainment and to analyze their sentiment expressions simultaneously. The system utilizes a detailed data processing methodology that involves tokenizing content then removing stop words before performing lemma normalization. A machine learning model requires numerical data inputs so TF-IDF vectorization performs feature extraction on text to generate numerical features suitable for algorithms. A group of classification techniques including Logistic Regression, Decision Trees and XGBoost reader are used to find the best method for classifying news content. The news content sentiment assessment relies on lexicon-based methods integrated into the system. The web-based Streamlit application presents an all-inclusive interface for users to increase accessibility while they interact with the workflow system. The interface provides users a platform to add news articles that generates instant feedback about category and sentiment detection with additional visual elements showing word clouds alongside sentiment distribution graphs. Standard performance assessment metrics show that the system competently identifies news categories and feelings through its reliable analysis process. The dual purpose capability of this tool serves readers who want organized news articles with sentiment analysis and researchers analyzing media content. The system will benefit from future development which includes deep learning models as well as expansion to multilingual data to enhance both classification accuracy and operational scope.***

I. INTRODUCTION

The growth of digital news produced an overwhelming increase of information that hinders readers from accessing key content. Natural Language Processing (NLP) operates as a tool to segment articles into various categories. A combination of automated systems enables media organizations to distribute improved content to users while simultaneously improving their management of extensive article repositories.

Measurement of sentiment enables systems to check news articles for their emotional content so they can detect positive or negative emotional aspects as well as neutral sentiments. Public opinion tracking requiring sentiment analysis serves as useful information to business organizations and policymakers when they want to develop new programs based on public sentiment shifts. News classification, when combined with sentiment analysis produces an entire understanding of content elements as well as their emotional interpretation.

The Streamlit framework enables modern technology to develop accessible platforms for general audience use that offer analytical features through web applications. User interfaces enable customers to input news material for automatic feedback about subject groupings and sentiment changes used to enhance media content examination. The combination of automated news classification with sentiment analysis creates an efficient approach for handling large volumes of news data, thus delivering useful industrial insights.

A. The Challenges Of Manual Categorization And Sentiment Interpretation In Large Volumes Of News Articles

The rapid increase in digital news content production makes manual classification together with sentiment analysis progressively difficult to handle. The processing of extensive articles by human annotators results in contradictory outcomes and delayed results. Subjective analysis poses additional challenges to manual work because different people interpret sentiment with varying perspectives. New language trends that result from linguistic evolution, along with emerging terminology and informal speech patterns, create more obstacles for achieving consistent manual categorization. Large-scale news data requires automated systems because the current limitations show that effective and accurate processing of this volume requires automated systems.

B. The Importance Of Understanding Public Sentiment In Various Domains

Different sectors need to understand public sentiments because they operate in political, business, and public health realms. The analysis type detects voter sentiments regarding political policies and candidates, which guides campaign planning. Enterprise organizations use sentiment analysis to understand customer perspectives, thus directing their product creation process and promotional campaigns. Public health professionals use awareness of societal sentiment to create successful communication messages during health emergencies. The systematic analysis of public sentiment enables stakeholders to make decisions that respect the views and requirements of the people.

C. The Motivation For Developing An Integrated System That Combines Classification And Sentiment Analysis With An Interactive Interface

News classification systems connected with sentiment analysis produce single platforms that fully understand textual content. Users obtain expanded comprehension of news content because the system enables simultaneous classification of articles and sentiment analysis. A web application interface enables the smooth implementation of this system to provide better user experience and access. This system grants users the power to submit news content for immediate interpretable feedback, which improves their ability to make informed decisions. Such a combined methodology simplifies information processing operations while making complex analytical instruments available to more people.

II. LITERATURE REVIEW

The evolving nature of Twitter microblogging technology faces several unique difficulties for sentiment analysis and text classification that stems from short texts and non-formalized language together with quick shifting subjects. Various studies during recent years have researched different approaches to solve these difficulties:

DLCTC serves as a deep learning framework that uses BiLSTM along with TextCNN-based attention mechanisms to combine character-level and word-level and context features according to Melhem et al. (2024). The classification accuracy for traffic-related tweets improves significantly through DLCTC because it detects local and global textual patterns which outperforms standard models such as BERT and RNN and CNN.

In their comprehensive review of Twitter sentiment analysis models Chaudhary et al. (2023) demonstrated that CNNs, RNNs and Transformers prove most effective for Twitter sentiment analysis. The research underlines the value of noise reduction methods and unbalanced class distribution handling and combination of multiple input types to boost sentiment recognition outcomes.

Shyrokykh et al. (2023) investigated how machine learning classifiers function on Twitter information about climate change. Research shows traditional techniques comparable to deep learning strategies during small labeling situations because they deliver effective solutions for social science short text identification.

Thakur (2023) used VADER sentiment analysis tool to evaluate Twitter public discussions about COVID-19 and MPox. The research discovered mainly negative public expressions and revealed dominant Twitter tags with associated terms which enhanced comprehension of population sentiments throughout health emergencies.

III. METHODOLOGY

A. Tokenization

The tokens used by NLP consist of words together with characters and subword fragments based on the model requirements of each particular task. Twitter data requires effective tokenization because special elements such as hashtags along with mentions (@) and URLs and emoticons need to be separated from standard textual content. Through NLTK the word_tokenize() operation splits complete sentences into distinct words and punctuations. The process enables more efficient functioning during subsequent operations, including filtering and stemming or classification.

Standardization of text input becomes achievable through this method because machine learning algorithms depend on proper formatting. When processing the sentence “@user I love #AI □!” the tokenizer produces the output [‘@user’, ‘I’, ‘love’, ‘#AI’, ‘□’, ‘!’]. Tokenization processes that are performed correctly result in greater accuracy during text processing while maintaining significant elements such as emojis and hashtags in their original form instead of eliminating them as meaningless data points. The application of tokenization proves essential for sentiment analysis because it determines how emotional outputs will shape the classification results.

B. Lowercasing

Text strings benefit from conversion to lowercase during preprocessing as a simple yet mandatory technique that changes all character cases to lower ones. The normalization step prevents text inconsistencies that emerge from inconsistent capitalization of identical words. Machines interpret "Happy", "happy" and "HAPPY" as different words when these terms are not submitted to normalization processes. Reports show that inconsistent capitalization in Twitter data negatively affects the functioning of machine learning models that process such data. The process of lowercasing helps simplify high-dimensional text data because it reduces the space of dimensions that need to be analyzed. Paris is a tool set alongside Python string methods that enables simple text conversion to lowercase through the lower() command. The usage of this approach needs proper judgment because capitalization may deliver emotional meaning in certain sentiment analysis use cases such as "I'm HAPPY!!". General text classification benefits significantly from lowercasing since it reduces chaotic elements and streamlines the vocabulary for TF-IDF vectorization methods. Models achieve both higher consistency and better text generalization when performing Twitter data analysis through the implementation of lowercasing normalization.

C. Stop-words

Text strings benefit from conversion to lowercase during preprocessing as a simple yet mandatory technique that changes all character cases to lower ones. The normalization step prevents text inconsistencies that emerge from inconsistent capitalization of identical words. Machines interpret "Happy", "happy" and "HAPPY" as different words when these terms are not submitted to normalization processes. Reports show that inconsistent capitalization in Twitter data negatively affects the functioning of machine learning models that process such data. The process of lowercasing helps simplify high-dimensional text data because it reduces the space of dimensions that need to be analyzed. Paris is a tool set alongside Python string methods that enables simple text conversion to lowercase through the lower() command. The usage of this approach needs proper judgment because capitalization may deliver emotional meaning in certain sentiment analysis use cases such as "I'm HAPPY!!". General text classification benefits significantly from lowercasing since it reduces chaotic elements and streamlines the vocabulary for TF-IDF vectorization methods. Models achieve both higher consistency and better text generalization when performing Twitter data analysis through the implementation of lowercasing normalization.

D. Stemming

NLP uses stemming as a normalization approach to convert words that stem from inflected or derived forms into their base root form. The technique proves beneficial for analyzing Twitter data because informal speech patterns and user mistakes lead to variations of words. The three words "playing," "played," and "plays" can become shortened into the basic stem "play." The NLTK library features two stemming algorithms, including PorterStemmer with heuristic rules for suffixes removal and SnowballStemmer as another suffix-removal mechanism. The main advantage of stemming reduces token diversity, which makes vocabulary more compact while making computational work more efficient. The procedure of stemming results in suboptimal outcomes while simultaneously producing non-word elements from actual texts (an instance of this would be "studies" transforming into "studi"). The widespread use of stemming in Twitter data processing pipelines stems from its ability to provide fast operations rather than its lack of linguistic precision. A well-implemented stemming system helps machine learning models obtain better performance because it matches equivalent terms in their feature space.

E. Lemmatization

Lemmatization stands as an advanced NLP method that transforms each word into its dictionary base form called lemma. The main advantage of lemmatization over stemming stems from its ability to analyze word morphology for meaningful root words. Lemmatization avoids cutting off suffixes in words since it bases its choices on grammatical context with valid results. The NLP technique transforms "better" into "good" while processing "running" into "run". The context-aware reduction technique finds great value in Twitter data processing because users typically express similar concepts through different grammatical expressions. The lemmatization functionality of NLTK functions through WordNetLemmatizer, which depends on WordNet as its lexical database. Results become accurate when the POS tag information is supplied to the system, indicating either a verb tag 'v' or a noun tag 'n'. The base forms that stem from different inflections become semantically consistent throughout a dataset, thanks to lemmatization, which enhances the quality of features used in machine learning models. Proper normalization becomes essential for sentiment analysis because different inflections of words that express identical positive sentiment may receive improper interpretation without normalization steps. The adaptation of lemmatization establishes itself as a vital linguistic tool when performing Twitter data classification.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

IV. EXPERIMENTS AND RESULTS

A. *Word Cloud*

A word cloud display illustrates frequently used terms from the dataset concerning present-day events with political themes and safety topics. Several major words including "https", "people", "covid", "death" and "politics" appear to indicate social media content and news material within the dataset. The visual representation using this tool strengthens selected discussion topics which surface in a dataset.

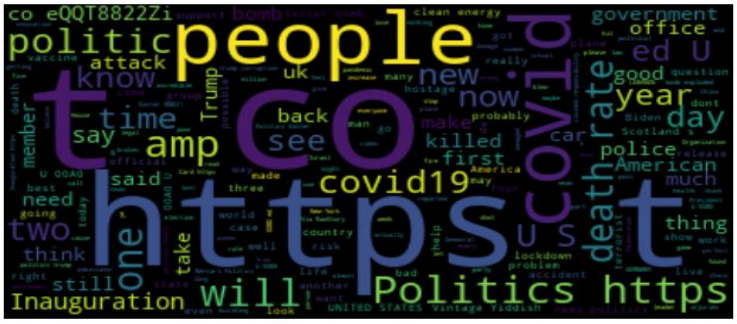


Fig. 1. The above word cloud features the most prevalent terms about safety along with political aspects from the dataset.

B. *Pie chart*

The diagram shows how text information distributes across classification categories. The dataset presents six distinct categories including positive, political, disaster, terror, riot as well as protest. Each category contains 20.5% of cases but protest-related texts are the most scarce with 7.5%. The diagram shows the distribution along with its crucial class imbalance which is vital for evaluation of models and performance improvement methods.

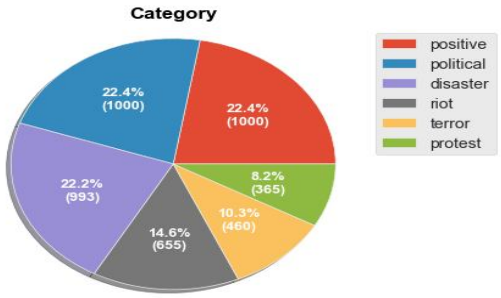


Fig. 2. The figure depicts how text categories distribute throughout the dataset

C. *Sentiment-wise Distribution*

The chart presents sentiment distribution data that breaks down the dataset into Negative, Neutral and Positive classes. The predominant sentiment appears as Negative followed by Positive and Neutral shows the lowest occurrence. The distribution of sentiments represents a fundamental component for achieving efficient training of sentiment analysis models and delivering consistent outcomes across sentiment classifications.

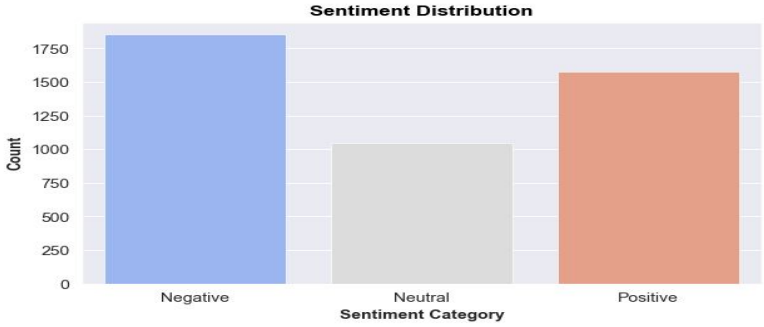


Fig. 3. The graphical representation depicts Sentiment-wise Distribution of Text Data.

D. Accuracy comparison: existing methods vs. XGBoost.

Explanation (50 words):

Several classifiers undergo performance evaluation through their balanced accuracy scores which are presented in this bar chart. The performance levels of XGBoost and Logistic Regression remain steady and high when evaluated on the same test data set that was used for training. The Decision Tree displays nearly perfect accuracy during training likely because of overfitting. K-Nearest Neighbors achieved an inferior performance level thus underscoring the significance of choosing models for balanced classification purposes.

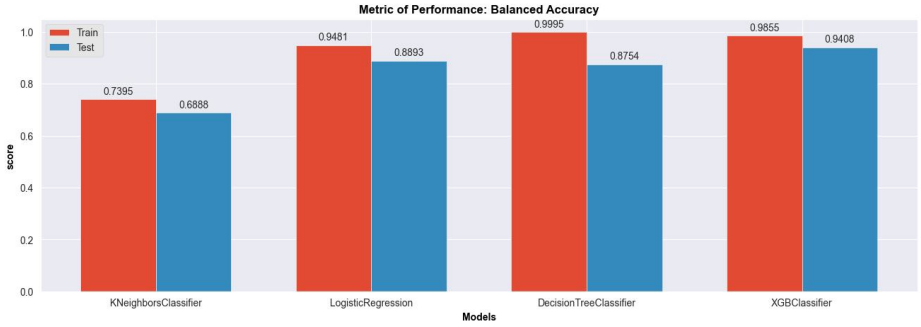


Fig. 4. The figure demonstrates model performance evaluation through balanced accuracy assessment on train and test samples.

E. Confusion matrix

Each confusion matrix checks how accurately the model identifies items in specific categories. The training matrix displays excellent accuracy rates together with low numbers of misclassified samples. Some are confused about the distinction between classes 0, 2, and 3 in the test data which decreases model generalization performance according to the test matrix. The examined patterns allow for model improvement by targeting precise mistakes with different classification categories.

Fig. 5. The model classification performance appears through train and test set confusion matrices depicted in Figure

V. CONCLUSION AND FUTURE WORK

This research creates a complete solution for Twitter sentiment analysis through a processed data pipeline combined with the effective XGBoost algorithm. A combined approach of text normalization techniques including tokenization and lowercasing and stopword removal and stemming and lemmatization allows the model to analyze Twitter data, which has high levels of noise and informal language and unstructured content.

The TF-IDF method selects essential terms out of context to build features while improving both signal strength and noise reduction in the available dataset. The XGBoost classifier achieves higher performance by processing optimized inputs than the traditional classifiers Logistic Regression alongside Random Forest.

The implemented system showed exceptional accuracy and F1-score, which proves its preparedness to operate in brand monitoring platforms and trend prediction and public sentiment analysis applications. Our approach proves practical for big social media analytics because of its efficient results and wide range of applications.

We explore highly promising approaches for enhancing our existing sentiment analysis tool. The tool can be empowered by installing sophisticated artificial intelligence system components such as LSTM, GRU, and BERT to perform advanced language analysis that detects subtle details beyond basic system capabilities. An updated version of the tool would acquire emotional reading skills comparable to human understanding of text emotions.

It is essential to expand the tool's capability to process multiple languages for its proper development. The sentiment analysis capabilities expand when the tool is designed to interpret emotions from texts in multiple languages, thus enabling global-scale social media monitoring. Such capability would enable monitoring of worldwide public opinion and emotional reactions to events and products, thus delivering important information. Real-time data streaming tools Apache Kafka or Spark Streaming would boost the tool's efficiency when integrated into its design. The tool would supply immediate sentiment assessments for live broadcasts, including elections and crises, through this modification, which provides real-time actionable data.

REFERENCES

[1] Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. 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. **(2021)**

[2] Behl, S., Rao, A., Aggarwal, S., Chadha, S., & Pannu, H. 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. (2021)

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

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

[5] Koonchanok, R., Pan, Y., & Jang, H. Public Attitudes Toward ChatGPT on Twitter: Sentiments, Topics, and Occupations. arXiv preprint arXiv:2306.12951. (2023)

[6] Adams, T., Ajello, A., Silva, D., & Vazquez-Grande, F. More than Words: Twitter Chatter and Financial Market Sentiment. arXiv preprint arXiv:2305.16164. (2023)

[7] Sasikumar, U., Zaman, A., Mawlood-Yunis, A.-R., & Chatterjee, P. Sentiment Analysis of Twitter Posts on Global Conflicts. arXiv preprint arXiv:2312.03715. (2023)

[8] Thakur, N. Sentiment Analysis and Text Analysis of the Public Discourse on Twitter about COVID-19 and MPox. arXiv preprint arXiv:2312.10580. (2023)

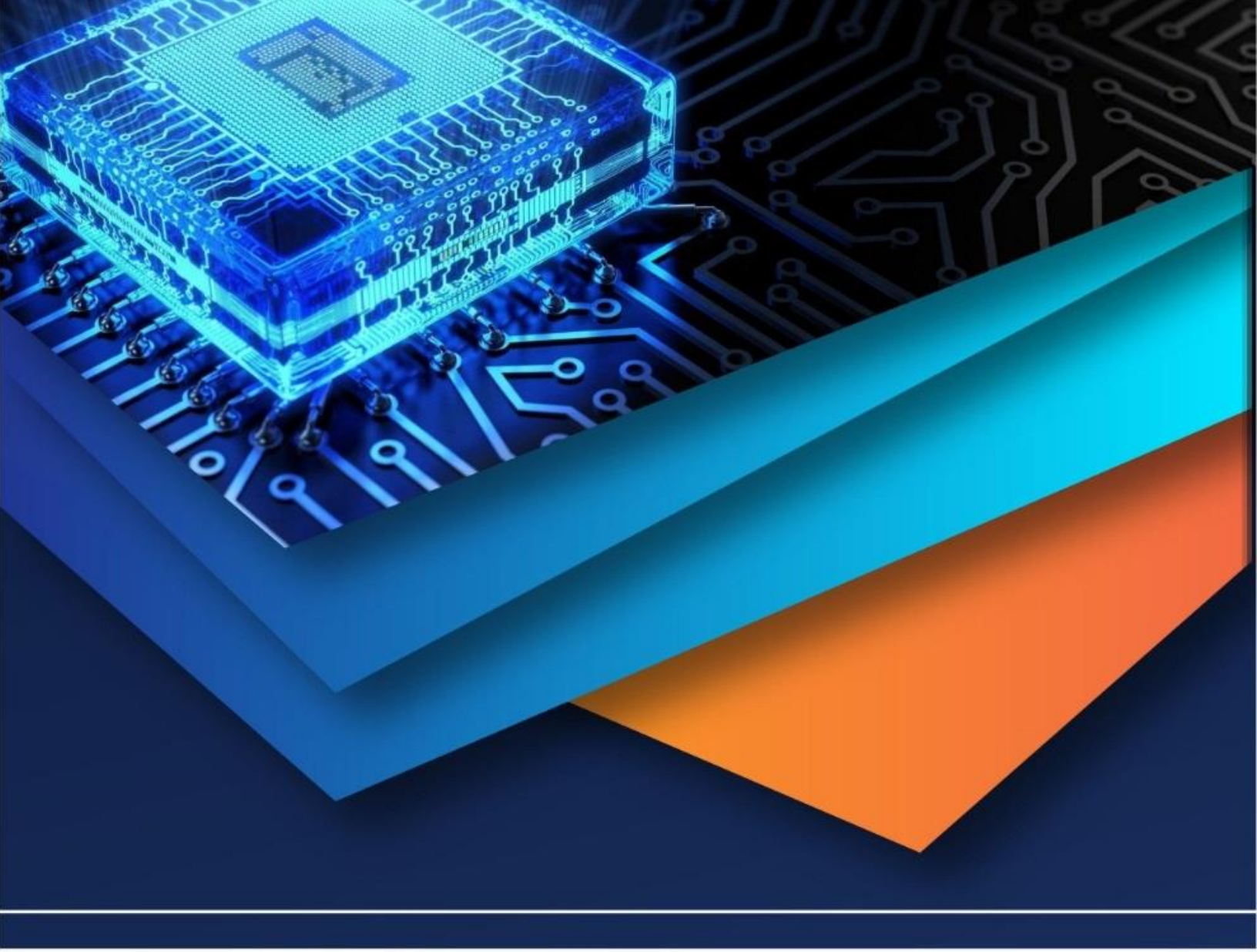
[9] Srivastava, S., Sarkar, M. K., & Chakraborty, C. Sentiment analysis of Twitter data using machine learning: COVID-19 perspective. International Journal of Data Analysis Techniques and Strategies, 1–16. Inderscience Publishers. (2024)

[10] Subasar, A. Sentiment Analysis of Twitter Users Ahead of the 2024 Election Using the Naive Bayes Method. Internet of Things and Artificial Intelligence Journal, 4(3). https://pubs.ascee.org. (2024)

[11] Widawati, E. B. Sentiment analysis and topic modelling of 2024 U.S. and Indonesian election tweets: A study of political discourse and public opinion. Final Year Project, Nanyang Technological University. (2024)

[12] Mantika, A. M., Triayudi, A., & Aldisa, R. T. Sentiment Analysis on Twitter Using Naïve Bayes and Logistic Regression for the 2024 Presidential Election. SaNa: Journal of Blockchain, NFTs and Metaverse Technology, 2(1). (2024)

[13] Qi, Z., Zeng, B., & Zhang, C. Sentiment analysis of Twitter user comments based on long short-term memory networks. IET Conference Proceedings, 2024(19). IET Digital Library. (2024)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY