iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83933

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 14 Issue VI June 2026- Available at www.ijraset.com

Neural Network Based Cyberbullying Classification

Amena Arman[1], Dr. P. Sammulal[2]
[1]*M. Tech Student Department of Computer Science and Engineering, JNTU Hyderabad, Telangana*
[2]*Professor of Computer Science and Engineering, JNTU Hyderabad, Telangana*

Abstract: ***The rapid growth of social media platforms has led to a significant increase in cyberbullying, making it a major challenge for online communities. Cyberbullying includes harmful and offensive behavior targeting individuals based on factors such as age, gender, religion, and ethnicity. Automatic detection of such content remains difficult due to informal language, sarcasm, ambiguous expressions, and varying contextual meanings. Therefore, there is a need for intelligent systems capable of accurately identifying different forms of cyberbullying. This research focuses on multiclass cyberbullying detection using advanced Natural Language Processing (NLP) and Deep Learning techniques. The proposed framework classifies social media text into categories including age, gender, religion, ethnicity, other cyberbullying, and not cyberbullying.***

To achieve effective classification, Bidirectional Encoder Representations from Transformers (BERT) is employed to capture contextual and semantic information from text. While existing machine learning and deep learning approaches have shown promising performance, but they face challenges in handling uncertain or ambiguous content. To address this limitation, the proposed framework integrates BERT with Neutrosophic Logic. BERT generates meaningful contextual representations, while Neutrosophic Logic utilizes Truth, Indeterminacy, and Falsity measures to model uncertainty during classification. This integration enhances the reliability and interpretability of multiclass cyberbullying detection, making the framework more effective for analyzing complex social media content.

Keywords: Cyberbullying, Social Media, Natural Language Processing, Deep Learning, BERT, Neutrosophic Logic, Multiclass Cyberbullying Detection, Classification.

I. INTRODUCTION

Social media platforms have transformed the way people communicate, allowing users to share information, opinions, and experiences instantly. Despite their benefits, these platforms have also contributed to the rise of cyberbullying, a form of online harassment that includes offensive, abusive, threatening, or hateful behavior directed toward individuals or groups. Cyberbullying can have serious psychological and emotional consequences, including anxiety, depression, and social isolation. As social media usage continues to grow, detecting and preventing cyberbullying has become an important challenge for ensuring safer online environments. Cyberbullying detection is difficult because social media content often contains informal language, slang, abbreviations, sarcasm, emojis, and context-dependent expressions. Traditional text classification techniques frequently struggle to accurately interpret such complex and ambiguous content. Various Machine Learning and Deep Learning approaches, including SVM, CNN, LSTM, BiLSTM, GRU, and MLP, have been applied to cyberbullying detection with promising results. However, these methods mainly depend on probability-based predictions and may not effectively address uncertainty in real-world social media data. To address these limitations, the proposed framework combines Bidirectional Encoder Representations from Transformers (BERT) with Neutrosophic Logic for multiclass cyberbullying detection. The system is trained on a Twitter dataset containing categories such as age, gender, religion, ethnicity, other cyberbullying, and non-cyberbullying. BERT is used to capture contextual and semantic information from text, while Neutrosophic Logic incorporates Truth, Indeterminacy, and Falsity measures to model uncertainty in classification decisions. This integration improves the reliability and interpretability of cyberbullying detection, making the framework more effective for analyzing complex social media content.

II. RELATED WORK

Cyberbullying has emerged as a major challenge with the rapid growth of social media platforms, leading researchers to develop various automated detection techniques. Early studies primarily relied on traditional machine learning algorithms such as Decision Tree, Random Forest, Support Vector Machine (SVM), Logistic Regression, and XGBoost for identifying different forms of cyberbullying.

Alqahtani and Ilyas [1] proposed an ensemble-based multiclass cyberbullying detection framework that combined multiple classifiers to improve classification performance across categories such as age, religion, ethnicity, and gender. Similarly, Balaji et al. [2] and Hzami et al. [3] demonstrated that machine learning models can effectively detect cyberbullying when combined with feature extraction methods such as TF-IDF and N-grams. However, these approaches depend heavily on handcrafted features and often struggle to capture contextual meaning and semantic relationships within text.

To overcome the limitations of traditional machine learning methods, researchers introduced deep learning techniques capable of automatically learning meaningful representations from textual data. Islam et al. [4] developed a multi-label cyberbullying detection framework using LSTM, BiLSTM, CLSTM, and BiGRU architectures, which effectively captured long-range contextual dependencies. Likewise, Alinejad [5] investigated Neural Networks, CNN, LSTM, and GRU models for multiclass cyberbullying detection and reported improved performance in identifying victim-specific cyberbullying categories. Balaji et al. [2] further proposed a hybrid CNN-BiLSTM architecture that combined local feature extraction with contextual sequence learning, resulting in better classification performance than individual deep learning models.

Recent advancements in Natural Language Processing have shifted attention toward transformer-based architectures, particularly BERT (Bidirectional Encoder Representations from Transformers). BERT has shown remarkable success in text classification tasks because of its ability to understand contextual information from both preceding and succeeding words. Farasalsabila et al. [6] applied BERT for multi-label cyberbullying detection and demonstrated superior performance compared to conventional machine learning and deep learning approaches. Similarly, Hzami et al. [3] reported that BERT consistently outperformed other classification models due to its ability to capture semantic relationships and contextual meanings within social media text. Despite their strong classification capabilities, transformer-based models mainly focus on prediction accuracy and often do not explicitly address uncertainty in classification outcomes.

Researchers have also explored hybrid and hierarchical frameworks to further improve cyberbullying detection performance. Hybrid approaches combine the strengths of different architectures to generate richer feature representations and improve classification accuracy. Balaji et al. [2] demonstrated that combining CNN and BiLSTM architectures enhances the model's ability to capture both syntactic and contextual information. In addition, Hzami et al. [3] proposed a multi-level cyberbullying detection framework in which binary classification is first performed to identify cyberbullying content, followed by multiclass classification to determine the specific cyberbullying category. Such hierarchical approaches simplify the classification process and improve overall detection effectiveness. Although advanced classification models have improved cyberbullying detection, handling uncertainty and ambiguity remains a significant challenge. Real-world social media posts often contain overlapping meanings, sarcasm, and vague expressions that make classification difficult. To address this issue, neutrosophic theory has been introduced as an uncertainty-handling mechanism. Vadivel et al. [7] demonstrated the effectiveness of neutrosophic score functions in decision-making under uncertain conditions, while Fan et al. [8] highlighted the advantages of neutrosophic-set-theory-based methods in representing incomplete and ambiguous information. These studies established a strong theoretical foundation for incorporating uncertainty management into text classification systems. Building upon these concepts, Ibrahim et al. [9] proposed a cyberbullying-related hate speech detection framework that integrated Neutrosophic Logic with a neural network classifier. Their approach utilized Truth, Indeterminacy, and Falsity values to represent classification confidence and uncertainty, resulting in improved fine-grained cyberbullying detection. However, most existing transformer-based studies focus primarily on classification accuracy, while neutrosophic-based cyberbullying approaches mainly rely on conventional neural networks. Therefore, a research gap exists in combining the contextual understanding capability of BERT with the uncertainty-handling strength of Neutrosophic Logic. The proposed work addresses this gap by integrating BERT and Neutrosophic Logic for multiclass cyberbullying classification, aiming to provide more accurate, reliable, and interpretable detection of cyberbullying content.

III. METHODOLOGY

A. System Overview

The proposed framework is designed to perform multiclass cyberbullying detection from social media text using BERT and Neutrosophic Logic. The system analyzes textual content collected from Twitter dataset and classifies it into six categories: Age, Gender, Religion, Ethnicity, Other Cyberbullying, and Not Cyberbullying. BERT is utilized to capture contextual and semantic information from the text, while Neutrosophic Logic is employed as a decision-making layer to handle uncertainty and ambiguity in classification results. The framework aims to provide accurate, reliable, and interpretable cyberbullying detection by combining contextual language understanding with uncertainty-aware reasoning.

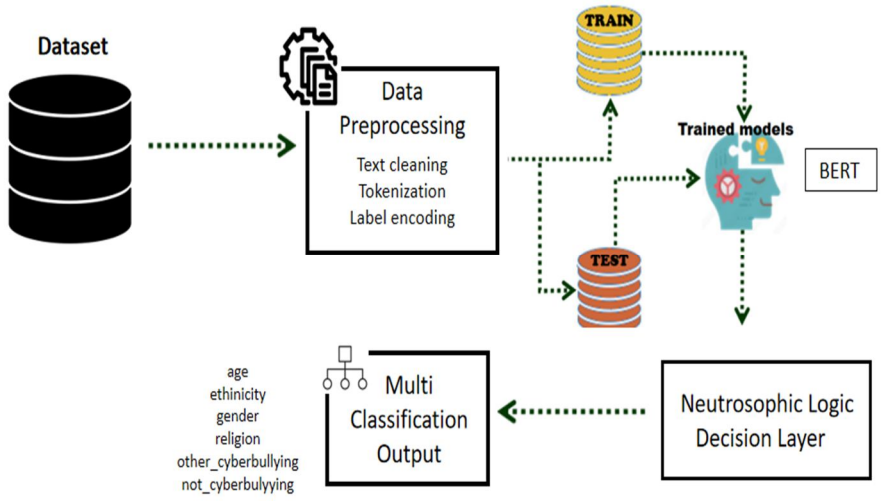


Figure 3.1: System Architecture

Figure 3.1 illustrates the overall architecture of the proposed Multiclass Cyberbullying Detection System. The framework begins with dataset collection and preprocessing, followed by training and testing of the BERT classification model. The generated probability scores are evaluated using the Neutrosophic Logic Decision Layer to handle uncertain predictions and improve classification reliability. The final output is a multiclass prediction that identifies the specific type of cyberbullying present in the social media text.

B. Data Acquisition

The dataset used in this research consists of social media comments collected from Twitter/X platforms. The dataset contains 47,692 labeled text samples categorized into six classes: age, gender, religion, ethnicity, other cyberbullying, and not cyberbullying. The dataset was selected because it contains diverse cyberbullying content commonly observed on social media platforms, enabling the model to learn various forms of abusive and harmful language. The balanced distribution of categories supports effective multiclass classification.

C. Preprocessing

Data preprocessing is performed to improve the quality and consistency of social media text before it is provided to the BERT classification model. Initially, the dataset is organized into six predefined cyberbullying categories, with each record containing textual content and its corresponding class label. The text then undergoes cleaning operations where URLs, hashtags, user mentions, emojis, punctuation marks, special characters, and unnecessary spaces are removed, and all text is converted to lowercase format. To ensure balanced learning across all categories, undersampling is applied using the minimum class count, preventing bias toward majority classes. The categorical labels are subsequently transformed into numerical values through Label Encoding, enabling the model to process the target classes efficiently. Finally, the dataset is divided into training and testing subsets, allowing the model to learn cyberbullying patterns from the training data and evaluate its performance on previously unseen social media text.

D. BERT-Based Classification Model

The proposed framework utilizes Bidirectional Encoder Representations from Transformers (BERT) as the primary classification model. Initially, the BERT tokenizer converts text into token IDs and attention masks. These tokenized inputs are passed through multiple Transformer Encoder layers that learn contextual relationships among words. The final [CLS] token embedding serves as the feature representation of the entire sentence. This representation is forwarded to fully connected layers and a Softmax activation function, which generate probability scores for all cyberbullying categories. The bidirectional nature of BERT enables the model to capture semantic and contextual information effectively, improving multiclass cyberbullying classification.

E. Neutrosophic Logic Decision Layer

Although BERT generates probability scores for each category, uncertainty may exist when multiple classes receive similar confidence values. To address this limitation, Neutrosophic Logic is integrated as a decision-making layer. The probability scores are transformed into Truth (T), Indeterminacy (I), and Falsity (F) values. Truth indicates confidence in a class, Indeterminacy measures uncertainty, and Falsity represents the likelihood that a class is incorrect. The final prediction is selected based on the maximum Truth value and minimum Indeterminacy value. This additional reasoning mechanism improves the reliability, transparency, and interpretability of the classification process.

F. Topic Extraction Module

In addition to cyberbullying classification, the framework performs topic extraction using Latent Dirichlet Allocation (LDA). The extracted topics provide insights into the primary themes discussed in the input text. This module helps users better understand the context associated with the detected cyberbullying content and enhances the interpretability of the overall system.

IV. EXPERIMENTAL RESULTS

The experimental results demonstrate the effectiveness of the proposed Multiclass Cyberbullying Detection System using BERT and Neutrosophic Logic. The framework analyzes social media text and classifies it into six categories: Age, Gender, Religion, Ethnicity, Other Cyberbullying, and Not Cyberbullying. The results show that BERT successfully captures contextual information from social media content, while Neutrosophic Logic improves decision-making by handling uncertain and ambiguous predictions.

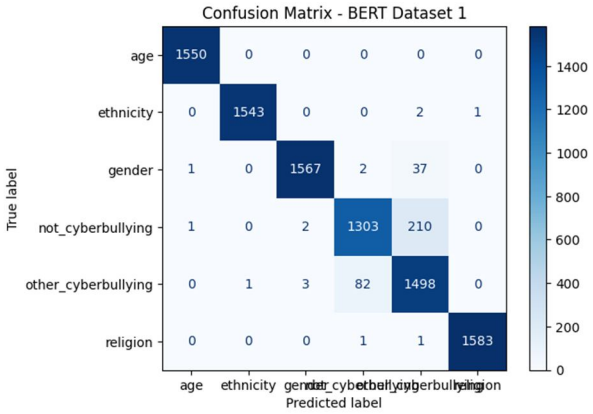


Figure 4.1: Confusion Matrix

Figure 4.1 presents the confusion matrix obtained from the BERT-based multiclass classification model. The matrix illustrates the relationship between actual and predicted cyberbullying categories before the application of the Neutrosophic Logic Decision Layer. Most samples are correctly classified into their respective categories, indicating the effectiveness of BERT in capturing contextual information from social media text. The generated probability scores are subsequently forwarded to the Neutrosophic Logic layer for uncertainty analysis and final decision support

Figure 4.2: Home Dashboard of the Cyberbullying Detection System

Figure 4.2 illustrates the main dashboard of the proposed Cyberbullying Detection System. The dashboard provides users with access to cyberbullying analysis features and serves as the primary interface for interacting with the application.

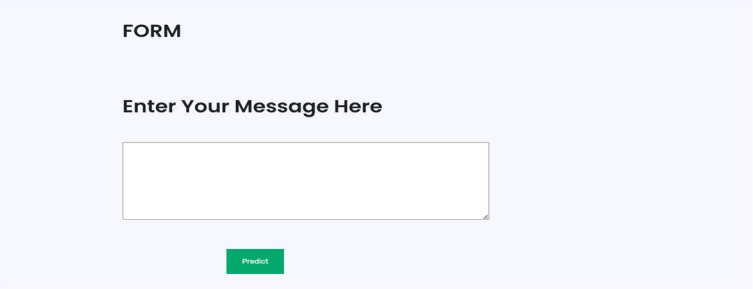


Figure 4.3: Text Input Interface

Figure 4.3 presents the text input interface where users can enter social media comments, tweets, or messages for cyberbullying analysis. This interface acts as the starting point of the prediction process.

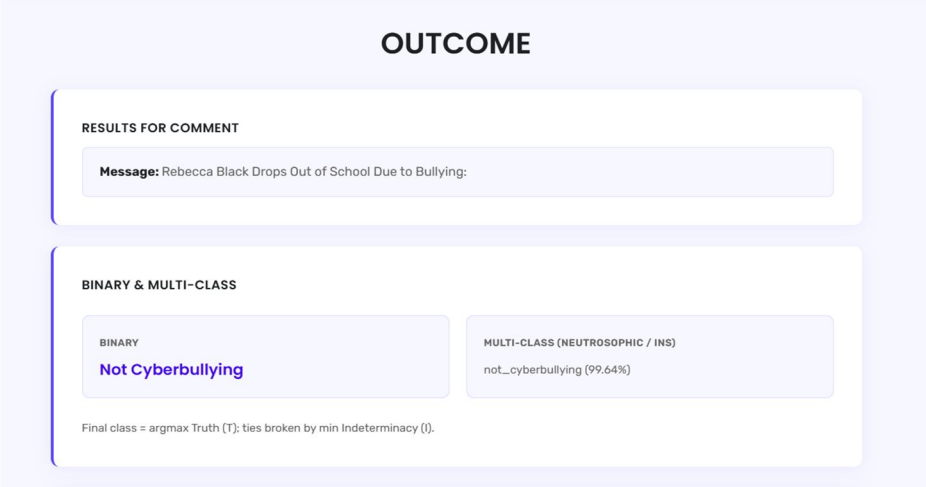


Figure 4.4: Cyberbullying Prediction Result

Figure 4.4 displays the classification result generated by the proposed framework. The system predicts the cyberbullying category and presents the corresponding confidence score. The output demonstrates the model's ability to classify social media text into one of the predefined categories.

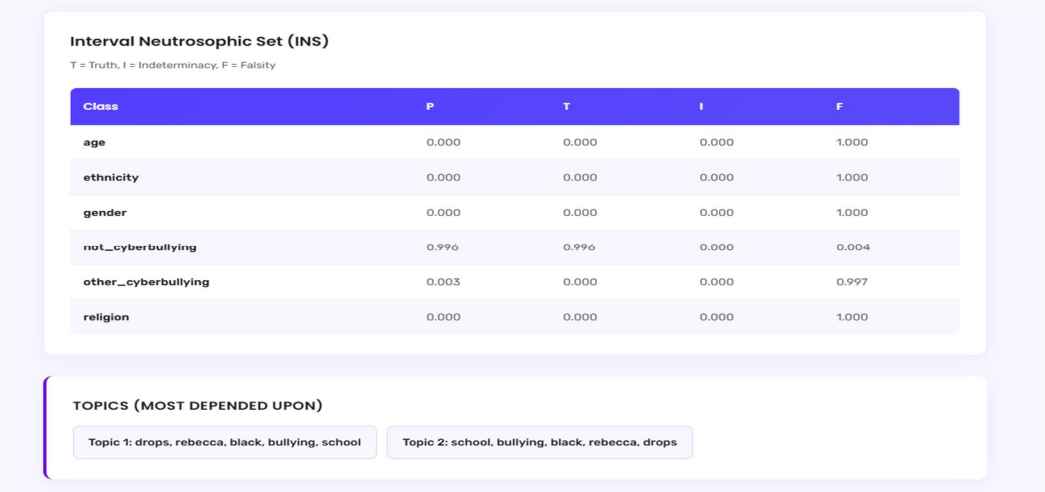


Figure 4.5: Neutrosophic Analysis Results

Figure 4.5 presents the Neutrosophic Logic analysis generated for the submitted text. The system computes Probability (P), Truth (T), Indeterminacy (I), and Falsity (F) values for each category. These values provide additional information regarding classification confidence and uncertainty, making the decision-making process more transparent and interpretable. The figure also displays the topics extracted from the input text using LDA, highlighting the key themes and contextual information associated with the detected cyberbullying content.

V. CONCLUSION

The proposed Multiclass Cyberbullying Detection System was developed to automatically identify different types of cyberbullying present in social media text using BERT and Neutrosophic Logic. The system effectively classifies textual content into categories such as Age, Ethnicity, Gender, Religion, Other Cyberbullying, and Not Cyberbullying. By leveraging BERT's contextual understanding capabilities, the model was able to capture meaningful linguistic patterns and semantic relationships within social media conversations, enabling accurate multiclass classification. To further enhance the reliability of predictions, Neutrosophic Logic was integrated into the framework as a decision-making layer. By incorporating Truth, Indeterminacy, and Falsity measures, the system was able to better handle uncertain and ambiguous cases, making the classification process more interpretable and trustworthy. In addition, the topic extraction module provided useful insights into the key themes discussed in the analyzed text, offering a deeper understanding of the content. The experimental results showed that the proposed framework effectively distinguished between cyberbullying and non-cyberbullying content while maintaining strong classification performance across multiple categories. The developed system demonstrates the potential of combining advanced Natural Language Processing techniques with uncertainty-aware reasoning to support the detection of harmful online behavior. Overall, the proposed approach provides a practical and intelligent solution for cyberbullying detection and can contribute to creating safer and more responsible online communication environments.

VI. FUTURE SCOPE

Future enhancements can focus on expanding the scope, accuracy, and practical applicability of cyberbullying detection systems. One important direction is the incorporation of multilingual support, which would enable the detection of cyberbullying across different languages and regions, making the framework more suitable for global social media platforms. The framework can also be extended to analyze multimedia content such as images, videos, and audio messages, allowing cyberbullying to be detected beyond textual communication. Furthermore, utilizing larger and more diverse datasets, along with advanced transformer-based models, may improve classification performance and strengthen the model's ability to handle real-world social media content. Another promising area is the integration of real-time social media monitoring, which would allow harmful content to be identified and flagged immediately after it is posted. Combining sentiment analysis and emotion detection techniques could provide deeper insights into user behavior and help assess the severity of cyberbullying incidents more effectively. In addition, Explainable Artificial Intelligence (XAI) techniques can be incorporated to provide clearer explanations for model predictions, improving transparency and user trust. With these advancements, the framework can evolve into a more comprehensive cyberbullying detection and monitoring solution capable of supporting safer and more responsible online communities.

REFERENCES

[1] A. F. Alqahtani and M. Ilyas, “An Ensemble-Based Multi-Classification Machine Learning Classifiers Approach to Detect Multiple Classes of Cyberbullying,” Machine Learning and Knowledge Extraction, vol. 6, no. 1, pp. 156–170, Jan. 2024, doi: 10.3390/make6010009.

[2] P. G. Balaji, P. P. Katariya, S. S., and M. Venugopalan, “Cyberbullying Detection on Multiclass Data Using Machine Learning and A Hybrid CNN-BiLSTM Architecture,” in Proc. 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), 2024, doi: 10.1109/ICKECS61492.2024.10616957.

[3] M. Hzami, H. Mahersia, and T. Bejaoui, “Multi-Level Cyberbullying Detection on Social Media Using Machine and Deep Learning Models,” in Proc. 2025 5th IEEE Middle East and North Africa Communications Conference (MENACOMM), 2025, doi: 10.1109/MENACOMM62946.2025.10911024.

[4] N. Islam, R. Haque, P. K. Pareek, M. B. Islam, I. H. Sajeeb, and M. H. Ratul, “Deep Learning for Multi-Labeled Cyberbully Detection: Enhancing Online Safety,” in Proc. 2023 International Conference on Data Science and Network Security (ICDSNS), 2023, doi: 10.1109/ICDSNS58469.2023.10245135.

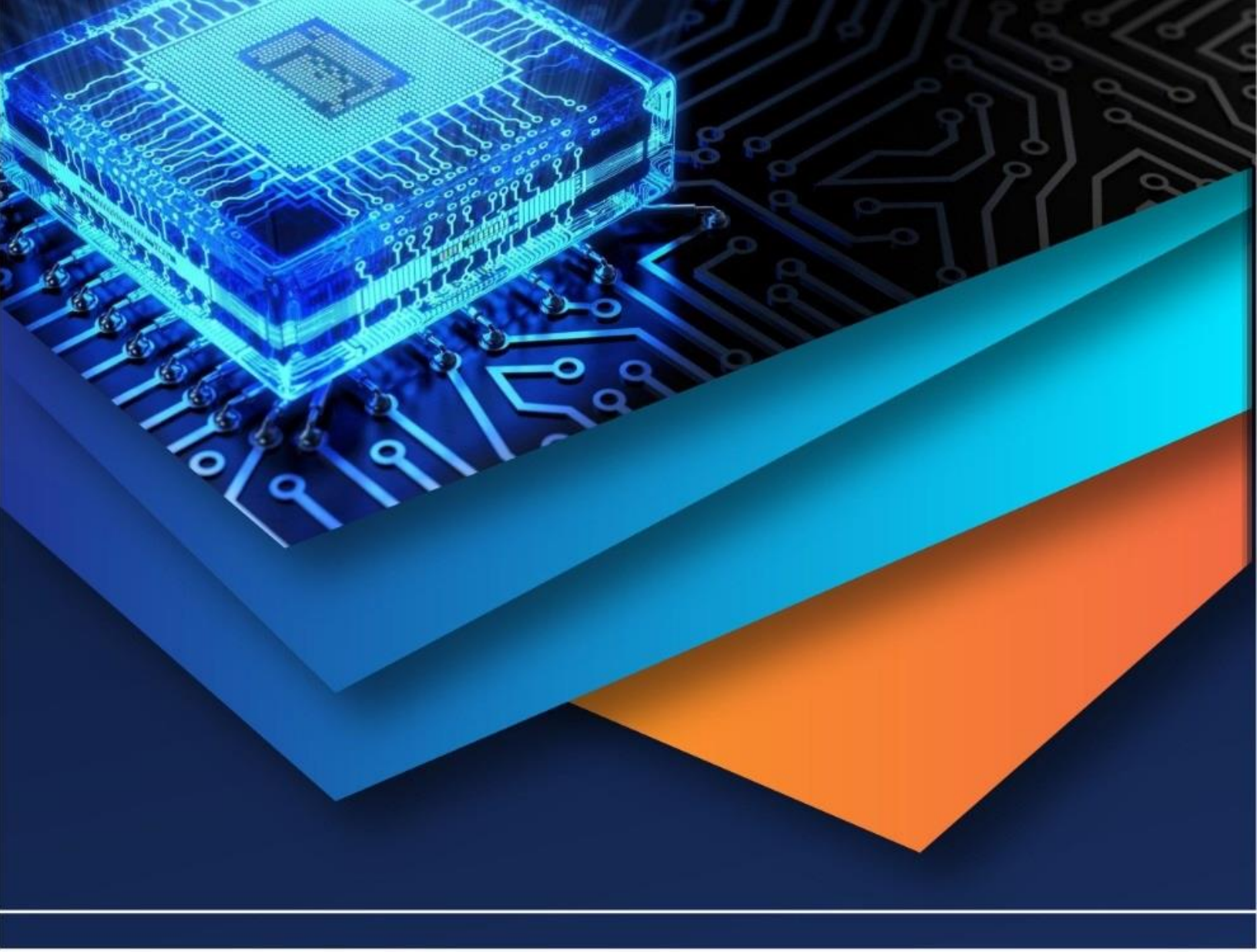
[5] M. Alinejad, “Multiclass Cyberbullying Detection Using Advanced Neural Network Architectures: A Comparative Study Amidst the COVID-19 Pandemic,” University of Central Florida STARS, Data Science and Data Mining, Jul. 2025. [Online]. Available:

[6] F. Farasalsabila, E. Utami, and S. Raharjo, “Multi-Label Classification Using BERT for Cyberbullying Detection,” in Proc. 2024 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA), 2024, doi: 10.1109/ICSINTESA62455.2024.10748045.

[7] A. Vadivel, N. Moogambigai, S. Tamilselvan, and P. Thangaraja, “Application of Neutrosophic Sets Based on Neutrosophic Score Function in Material Selection,” in Proc. 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2022, doi: 10.1109/ICEEICT53079.2022.9768448.

[8] E. Fan, K. Hu, and X. Li, “Review of Neutrosophic-Set-Theory-Based Multiple-Target Tracking Methods in Uncertain Situations,” in Proc. 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 2019, pp. 19–27.

[9] Y. M. Ibrahim, R. Essameldin, and S. M. Saad, “Social Media Forensics: An Adaptive Cyberbullying-Related Hate Speech Detection Approach Based on Neural Networks With Uncertainty,” IEEE Access, vol. 12, pp. 59474–59489, 2024, doi: 10.1109/ACCESS.2024.3393295.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY