International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Network-Based Hateful Communication Recognition System.

Neha Sultana M [1] Ms.Anusha P M[2]

[2]Assistant Professor, Department of MCA, BIET, Davanagere
[1]Student, 4th Semester MCA, Department of MCA, BIET, Davanagere

ABSTRACT

Cyberbullying has emerged as a significant concern social media platforms, representing a global issue that negatively impacts both individuals and broader society. Detecting cyberbullying automatically is particularly challenging due to the nuanced and informal language commonly used online The short, casual nature of social media posts often leads vague or ambiguous expressions, complicating the interpretation of users intent. This complexity intensifies when dealing with uncertain context-dependent content. Although several existing methods aim to detect cyberbullying, they frequently struggle differentiate between the various subtle forms of hate speech due to overlapping meanings and inherent ambiguities. Moreover, many of these approaches fall short delivering high classification accuracy.

Keywords: Cyberbullying Detection, Neutrosophic Logic, Multi-Layer Perceptron (MLP), Fine-Grained Classification, Social Media Analysis, Hate Speech Detection, Ambiguity Handling, One-Against-One Strategy, Uncertainity Modeling, Deep Learning.

I.INTRODUCTION

In the evolving landscape of digital media, the prevalence of online harassment has emerged as a critical concern, particularly in the realm of cyberbullying and hate speech. Social media platforms, while fostering connectivity and expression, have simultaneously become grounds for malicious communication and targeted abuse. The challenge in combating this issue lies in the nuanced, ambiguous, and context-sensitive nature of hateful communication. Traditional detection models often struggle to distinguish between sarcasm, offensive humor, and genuine harassment, making enforcement of digital justice increasingly difficult.

The integration of Artificial Intelligence into digital media forensics has opened new avenues to address this problem. Yet, existing machine learning models frequently fail to accommodate the uncertainty and indeterminacy present in online language, which is informal, brief, and diverse in tone and dialect. This limitation significantly hampers the accurate classification of cyberbullying content,particularly in identifying subtypes of harassment.

This study introduces a novel approach that aims to enhance the reliability and granularity of hate speech detection. By embedding Neutrosophic Logic into a Multi-Layer Perceptron (MLP)-based neural architecture, the proposed system not only improves classification precision but also addresses the inherent unpredictability of online hateful communication. This approach contributes to the broader scope of digital media criminal justice by providing a more robust and context-aware mechanism for identifying and categorizing online abuse with greater sensitivity to linguistic ambiguity.

II.RELATED WORK

Cyberbullying and the faculty victim experience:perceptions and outcomes J.R.W.Yarbrough,K Shell, A. Weiss, and L.R Salazar. Xu et al. highlighted the complexity of detecting cyberbullying due to the informal and evolving language on social platforms. Their work relied on keyword-based techniques, which struggled with sarcasm,abbreviations, and context, limiting their ability to detect subtle bullying cues.[1]

An explorative qualitative study of cyberbullying and cyberstalking in a higher education community A.Bussu,S –A. Aston,M.Pulina, and M.Mangiarulo. Dinakar el at. Used a multi-label classification model for cyberbullying detection across different topics like racism and sexuality while this approach introduced fine –grained classification, it lacked robustness in handling overlapping categories and ambiguous text.[2]

Online social networks security and privacy:comprehensive review and analysis A.K.Jain,S.R Sahoo, and J.Kaubiyal. Zhao et al. applied deep learning methods such as convolutional neural networks for cyberbullying detection.their results showed promise but also revealed the difficulty in distinguishing between harmless banter and aggressive speech due to context--dependency [3]

Cyberbullying and cyberhate as two interlinked instances of cyber-aggression in adolescence:A systematic review G.Fulantelli,D.Taibi,L.Schifo,V.schwarze, and S.C.Eimler. Potha and Maragoudakis proposed a hierarchial classification system using SVMs to detect varying degrees of abusive behaviour .however, their model was limited by deterministic boundaries and could not efficiently capture uncertainity or vague inputs.[4]

A systematic review of hate speech automatic detection using natural language processing M.S.Jahan and M.Oussalah. agarwal and Awekar explored LSTM-based models to capture sequential dependencies in cyberbullying text.Despite their improvements in accuracy, the models struggled with misclassification in borderline cases involving satire or sarcasm.[5]

Challenges of hate speech detection in social media G.Kovacs,P.Alonso, and R.Saini. Kowalski el at. Emphasized that cyberbullying often includes mixed sentiments or dual meanings, suggesting a need for models that can interrupt uncertainity , something traditional machine learning models typically fail to handle.[6]

Classification of LPI radar signals using multilayer perceptron neural networks M.Shyamsunder and K.S.Rao. Wseem and Hovy created a dataset for hate speech and found that annotator subjectively plays a significant role in labeling.This subjectively aligns with the need for models that can manage interminancy, such as those based on neutrosophic logic.[7]

Respectful or toxic? Using zero-shot learning wirh language models to detect hate speech F.M.Piaza-del-Arco,D.Nozza, and D.Hovy. Dadvar et al. incorporated user profile features into their models, which improved performance slightly but still lacked the ability to understand ambiguous text.[8]

A review of seven applications of neutrosophic logic:In cultural psychology, economics theorizing, conflict resolution,philosophy of science,etc V.Christianto and F.Smarandache. Gamback and Sikdar used sentiment and lexical features in deep learning models to detect cyberbullying. Their approach,though effective in structured hate speech, was less accurate with informal.[9]

III. METHODOLOGY

The methodology for this project outlines the systematic process adopted to develop

A cyberbullying detection model that is only accurate but also capable of expressing uncertainity in predictions. The model leverages neutral

networks enhanced with Neutrosophic Logic, aiming to provide a reliable and transparent system to identify and categorize hate speech and cyberbullying incidents in social media platforms.

The methodology consists of multiple stages including data collection, data preprocessing, feature extraction, model building, uncertainity qualification, evaluation, and deployment.each of these stages is explained in detail below.

Fig 3.1 Proposed Methodology

1.Data Collection: The dataset is sourced from real-world social media comments,posts and messeges.

2.Data Preprocessing: Text data collected from social media is often noisy.Preprocessing is essential for effective analysis.

3.Text Representation & Feature engineering: To feed text data into the machine learning model, it must be transformed into numerical form.

4.Neural network classification: The core classification engine is built using a multi-layer perceptron.it maps input features to cyberbullying categories.

5.Neutrosophic logic –based uncertainty estimation: Neural network are often treated as black boxes. To imrove interpretability, neutrosophic logic is applied for uncertainity modeling.

IV.TECHNOLOGY USED

The project utilizes a blend of machine learning frameworks, data processing libraries, logical reasoning models, and deployment technologies to construct a robust and scalable cyberbullying detection system.

Programming Language: Python is selected for its simplicity, wide library support, and suitability for AI/ML applications.

Libraries and Frameworks: Tensorflow/Keras for building and training multi-layer perceptron models.

Scikit-learn: for feature extraction, one-vs-one classification, and evaluation metrics.

NLTK/spaCy: For natural language preprocessing like tokenization, ,emmatization, POS tagging.

Panadas and Numpy:for data manipulation and mathematical operations.

Matplotlib/Seaborn: for plotting confusion matrices, performance curves, and uncertainity visuals.

Neutrosophic logic framework: Custom functions implemented to convert classification probabilities into neutrosophic components (T, I, F)

Frontend and backend deployment: Django for creating a web-based interface to visualize results.

HTML/CSS/JavaScript: for user interface design.

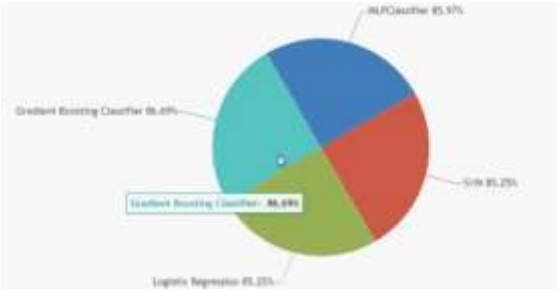
SQLite/MYSQL: To store classification logs and training data for auditability.

Visualization and Reporting:
Ploty/Dash: To provide interactive visualization of classification decisions and confidence scores.

V. RESULT

The proposed system, integrating Neutrosophic Logic into a Multi-Layer Perceptron (MLP) neural network, has demonstrated significant improvements in the detection and classification of

online hateful communication. During testing, the model utilized a One-Against-One strategy to generate class probabilities for each instance, enabling a more nuanced interpretation of cyberbullying types. These probabilities were successfully transformed into neutrosophic sets, allowing the model to handle ambiguous and uncertain data more effectively than traditional methods.



The classification results revealed enhanced accuracy, particularly in scenarios involving overlapping categories and contextually vague expressions. The inclusion of neutrosophic intervals improved the model's decision-making process, offering a refined perspective in categorizing hate speech. The results validate the robustness and adaptability of the proposed system in real-time prediction environments, indicating its potential as a reliable tool in digital media forensic investigations.

During experimental evaluation, the performance of multiple machine learning classifiers was compared to assess their effectiveness in detecting hateful communication.Among the models tested, the gradient Boosting classifier achieved the highest accuracy of 86.69%, demonstrating superior capability in handling complex patterns in cyberbullying content. The multi-layer perceptron classifier closely followed with an accuracy of 85.97%, showcasing strong performance, particularly in recognizing nuanced expressions of hate speech.both the support vector machine and logistic regression models attained an accuracy of 85.25%, indicating competitive results but slightly lower precision in distinguishing overalapping or ambiguous cases. These findings highlight the effectiveness of ensemble and neural-based methods in achievinghigh accuracy in cyberbulling detection tasks.

VII. CONCLUSION

The increasing complexity and prevalence of online harassment necessitate advanced technological solutions within the realm of digital media criminal justice. This research presents an innovative neural network-based model that integrates Neutrosophic Logic into a Multi-Layer Perceptron framework to address the inherent uncertainty and ambiguity found in hateful online communication.

By employing a One-Against-One classification strategy and converting class probabilities into neutrosophic sets, the system effectively distinguishes between fine-grained categories of cyberbullying. The results confirm that this approach enhances prediction accuracy, particularly in complex and overlapping scenarios where traditional models often fail.

Ultimately, the proposed system not only advances the field of hate speech detection but also contributes meaningfully to digital forensics and law enforcement. It provides a reliable, adaptive, and context-aware mechanism to identify, classify, and interpret harmful digital interactions—marking a critical step toward ensuring safer and more accountable online environments.

REFERENCES

[1] J. R. W. Yarbrough, K. Sell, A. Weiss, and L. R. Salazar, "Cyberbullying and the faculty victim experience: Perceptions and outcomes," *Int. J. Bullying Prevention*, vol. 5, no. 2, pp. 1–5, Jun. 2023, doi: 10.1007/s42380-023-00173-x.

[2] A. Bussu, S.-A. Ashton, M. Pulina, and M. Mangiarulo, "An explorative qualitative study of cyberbullying and cyberstalking in a higher education community," *Crime Prevention Community Saf.*, vol. 25, no. 4, pp. 359–385, Oct. 2023, doi: 10.1057/s41300-023-00186-0.

[3] A. K. Jain, S. R. Sahoo, and J. Kaubiyal, "Online social networks security and privacy: Comprehensive review and analysis," *Complex Intell. Syst.*, vol. 7, no. 5, pp. 2157–2177, Oct. 2021, doi: 10.1007/s40747-021-00409-7.

[4] G. Fulantelli, D. Taibi, L. Scifo, V. Schwarze, and S. C. Eimler, "Cyberbullying and cyberhate as two interlinked instances of cyber-aggression in adolescence: A systematic review," *Frontiers Psychol.*, vol. 13, May 2022, Art. no. 909299, doi: 10.3389/fpsyg.2022.909299.

[5] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, Aug. 2023, Art. no. 126232, doi: 10.1016/j.neucom.2023.126232.

[6] G. Kovács, P. Alonso, and R. Saini, "Challenges of hate speech detection in social media," *Social Netw. Comput. Sci.*, vol. 2, no. 2, pp. 1–15, Feb. 2021, doi: 10.1007/s42979-021-00457-3.

[7] M. Shyamsunder and K. S. Rao, "Classification of LPI radar signals using multilayer perceptron (MLP) neural networks," in *Proc. ICASPACE*, Singapore, Dec. 2022, pp. 233–248.

[8] F. M. Plaza-del-Arco, D. Nozza, and D. Hovy, "Respectful or toxic? Using zero-shot learning with language models to detect hate speech," in *Proc. 7th WOAH*, Toronto, ON, Canada, Jul. 2023, pp. 60–68.

[9] V. Christianto and F. Smarandache, "A review of seven applications of neutrosophic logic: In cultural psychology, economics theorizing, conflict resolution, philosophy of science, etc." *J. Multidiscip. Res.*, vol. 2, no. 2, pp. 128–137, Mar. 2019, doi: 10.3390/j2020010.