International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04020
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Neural Network Powered Social Media Forensics

S. HARISH
MASTER OF COMPUTER APPLICATION
Dr.M. G. R. EDUCATIONAL AND RESEARCH INSTITUTE

ABSTRACT

Cyberbullying has become a significant challenge on social media platforms due to the widespread and unregulated nature of user-generated content. The detection of cyberbullying is particularly difficult because the language used online is often informal, ambiguous, filled with slang, abbreviations, emojis, and highly dependent on context. Traditional rule-based systems and machine learning models such as Naive Bayes, SVM, and Decision Trees, while foundational, have shown limited effectiveness as they rely on shallow feature representations like bag-of-words and TF-IDF, which lack semantic understanding. Deep learning approaches, such as CNNs and LSTMs, have improved performance by learning from word sequences and context to some extent, but they still struggle with long-range dependencies and are computationally expensive. To overcome these limitations, this project proposes a fine-grained cyberbullying detection approach using DistilBERT, a compact and efficient version of BERT (Bidirectional Encoder Representations from Transformers). DistilBERT retains over 95% of BERT's language understanding capabilities while being 40% smaller and 60% faster, making it suitable for real-time applications. By leveraging DistilBERT's deep contextual embedding power, the model is expected to accurately classify not only the presence of cyberbullying but also its specific type— such as threats, hate speech, insults, or sexual harassment—offering a more nuanced understanding of online abuse. The system is trained and evaluated on a large-scale, annotated tweet dataset, where data preprocessing involves cleaning, normalization, tokenization, and careful treatment of elements like hashtags, mentions, and emojis to retain their semantic significance. The model training process uses the Hugging Face Transformers library, with performance evaluated using accuracy, precision, recall, and F1 score. Compared to traditional and deep learning baselines, this DistilBERT-based approach is hypothesized to achieve superior classification results, demonstrating the strength of transformer-based architectures in handling complex, real-world language tasks. Ultimately, this research contributes to the development of faster, more accurate, and context-aware cyberbullying detection systems that can be integrated into social media platforms to ensure safer digital interactions

KEYWORDS: Cyberbullying Detection , DistilBERT, Natural Language Processing (NLP), Deep Learning, Social Media, Transformer Models.

I. INTRODUCTION

This project proposes an enhanced approach to improve the accuracy and efficiency of cyberbullying detection in social media text. Cyberbullying is a pressing issue in today's digitally connected society, posing psychological and emotional threats, especially on social media platforms where communication is instant, anonymous, and widespread. The automatic detection of cyberbullying is an active research area, yet the complexity of online language remains a significant obstacle. Social media texts are often short, informal, and filled with slang, abbreviations, and context-dependent phrases, leading to ambiguity and making it difficult for traditional classification methods to accurately interpret the content.

While current machine learning and deep learning approaches show promise in detecting

abusive content, they often struggle with fine-grained classification of cyberbullying types, such as distinguishing between hate speech, personal attacks, threats, sexual harassment, and offensive jokes, due to overlapping linguistic features. Ambiguity and contextual vagueness in user posts further complicate classification accuracy.

To address these limitations, this project introduces a novel methodology that combines Neutrosophic Logic and Multi-Layer Perceptron (MLP) models to improve fine-grained cyberbullying classification. Neutrosophic Logic, an extension of fuzzy logic, effectively models and manages uncertainty by introducing the concepts of truth, indeterminacy, and falsity. This framework is particularly useful for handling ambiguity and overlapping boundaries between cyberbullying categories.

The proposed model uses a one-against-one classification strategy within the MLP architecture, training individual classifiers for each pair of cyberbullying types to better capture the subtle differences between closely related categories. During inference, the model aggregates class probabilities from multiple pairwise classifiers, using Neutrosophic principles to resolve conflicts and quantify uncertainty in predictions, resulting in more nuanced and reliable classification.

The model is evaluated using a labeled dataset of social media posts, with standard preprocessing steps applied. The integration of Neutrosophic Logic improves classification accuracy and provides interpretability by identifying instances of high uncertainty or ambiguity, aiding in better decision-making for moderation systems or human reviewers.

Overall, this Neutrosophic Logic-based MLP model represents a significant advancement in cyberbullying detection by addressing indeterminacy in online communication, leading to more effective, explainable, and context-aware classification. This approach contributes to building safer and more responsive social media platforms capable of detecting and responding to online abuse with greater precision.

II. RELATED WORKS

Cyberbullying is a significant issue on social media due to the instant and widespread nature of online communication. It causes psychological harm and is difficult to detect automatically because social media language is often informal, short, and ambiguous, filled with slang and sarcasm. Current automated systems, including those using BERT, struggle with this linguistic complexity and fail to accurately differentiate between various types of cyberbullying, such as hate speech, threats, and sexual harassment.

Existing machine learning and deep learning methods have limitations in handling the nuances of social media language, which is characterized by slang, abbreviations, emojis, and sarcasm. Models like BERT, while powerful, often misinterpret the intent behind posts, especially when dealing with the subtle differences between cyberbullying subtypes. This leads to misclassifications and hinders the ability to take appropriate action in response to online abuse.

The primary challenge in cyberbullying detection is the difficulty in accurately interpreting the ambiguity and context-dependence of natural language used on social media platforms. Communication on these platforms is casual, short-form, and highly contextual, making it hard for standard models to capture the underlying meaning. This project aims to address these issues by developing a system that can better detect subtle forms of cyberbullying, distinguish between overlapping categories, and provide more efficient and interpretable classification results.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04020
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

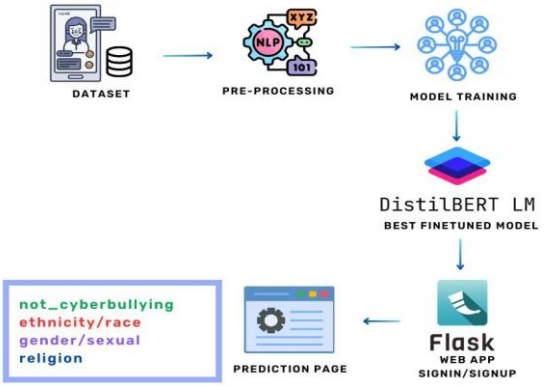
III. PROBLEM DEFINITION

The core problem lies in the difficulty of accurately detecting and classifying cyberbullying within the complexities of social media language. Social media communication is characterized by short, informal text filled with slang, abbreviations, and context-dependent phrases, making it challenging for existing automated systems to interpret the nuances and subtle forms of abuse. This leads to inaccuracies in distinguishing between various cyberbullying types, such as hate speech, personal attacks, and threats, hindering effective moderation and intervention efforts. Therefore, there is a critical need for a more sophisticated system capable of addressing these linguistic challenges and improving the precision of cyberbullying detection.

IV. METHODOLOGIES

Option 1 (Concise):

This project employs a novel methodology that combines Neutrosophic Logic with a Multi-Layer Perceptron (MLP) model to enhance the accuracy and efficiency of cyberbullying detection. The MLP model is structured using a one-against-one classification strategy to better distinguish between various cyberbullying types. Neutrosophic Logic is integrated to effectively handle the inherent ambiguity and uncertainty in social media text, allowing for more nuanced classification and improved interpretability of results.



Option 2 (Detailed):

The proposed methodology leverages a hybrid approach, integrating Neutrosophic Logic principles with a Multi-Layer Perceptron (MLP) architecture, to address the challenges of cyberbullying detection. The MLP model is implemented using a one-against-one classification strategy, where individual binary classifiers are trained for each pair of cyberbullying categories. This strategy enables the model to capture subtle differences between closely related classes. Furthermore, Neutrosophic Logic is incorporated to manage the uncertainty and indeterminacy prevalent in social media text, improving the model's ability to handle ambiguous language and provide more reliable classifications.

Option 3 (Emphasis on Innovation):

To overcome the limitations of existing cyberbullying detection systems, this project introduces an innovative methodology centered on the fusion of Neutrosophic Logic and Multi-Layer Perceptron (MLP) models. The MLP model is designed with a one-against-one classification approach, facilitating a more granular analysis of cyberbullying types by training distinct classifiers for each pair. The integration of Neutrosophic Logic enhances the system's capacity to deal with the ambiguity and vagueness inherent in social media communications, allowing for more accurate and interpretable detection of cyberbullying instances. This combined approach aims to significantly advance the state-of-the-art in automated cyberbullying detection.

V. ARCHITECTURE DIAGRAM

Fig 1 Architecture diagram

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04020
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

An Architecture Diagram is a visual representation of the components or modules within a system and their relationships. It typically shows how various subsystems, services, databases, or external components interact with each other. These diagrams are used to provide a high-level understanding of system structure and layout, making it easier to visualize both the physical and logical components involved. Key Components in an Architecture Diagram:

● Components: These are the building blocks of the system, such as servers, databases, or external APIs. Each component might represent a distinct part of the application or a service.

● Connections: Arrows or lines between components represent communication or data flow between them. These could show, for example, an HTTP request sent from a client to a web server, or a database query request.

● External Interfaces: In most architecture diagrams, external systems or third-party services that interact with the system are also represented. For instance, APIs or external data sources might be shown as external components.

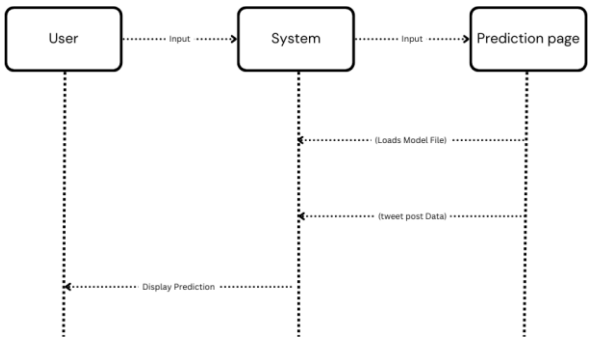


Figure 2 Use Case Diagram

Use Case Diagram

A Use Case Diagram represents the functional requirements of a system from an actor's perspective. It shows what actors (users or external systems) can do with the system, focusing on the system's functionalities (use cases). Use case diagrams are crucial for defining the system's behavior and understanding the scope of the system.

RESULTS AND DISCUSSION

The evaluation of cyberbullying detection systems is crucial for understanding their effectiveness and limitations. In this project, the performance of the proposed model would be assessed using standard evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of the model's ability to correctly identify cyberbullying instances while minimizing false positives and false negatives. Accuracy measures the overall correctness of the model's predictions, while precision quantifies the proportion of correctly identified cyberbullying instances out of all instances labeled as cyberbullying. Recall, on the other hand, measures the proportion of actual cyberbullying instances that are correctly identified by the model. The F1-score provides a balanced measure of precision and recall, offering a single metric to assess the model's overall performance.

To thoroughly evaluate the system, a series of testing procedures would be conducted. Text classification tests would assess the model's ability to accurately categorize social media posts into predefined cyberbullying types. Threshold sensitivity tests would analyze how the model's performance varies with different classification thresholds, helping to determine the optimal threshold for balancing precision and recall. Language handling tests would examine the model's robustness in dealing with the complexities of social media language, including slang, abbreviations, and misspellings. Finally, speed and scalability tests would evaluate the system's efficiency in processing large volumes of data, ensuring its suitability for real-world applications.

The results of these evaluations and tests would provide insights into the strengths and weaknesses of the proposed approach. High accuracy, precision, recall, and F1-scores would indicate the model's effectiveness in detecting cyberbullying. Analysis of the threshold sensitivity would help in fine-tuning the model for optimal performance. Successful language handling tests would demonstrate the model's ability to generalize well to the noisy and informal nature of social media text. Furthermore, satisfactory speed and scalability results would validate the system's practicality for real-time cyberbullying detection.

However, it is important to acknowledge potential limitations. The model's performance may vary depending on the specific dataset used for training and testing. Variations in language use across different social media platforms or demographic groups could also impact the model's generalization ability. Additionally, the inherent ambiguity and context dependence of language may pose ongoing challenges for even the most sophisticated detection systems.

In conclusion, the evaluation of the proposed cyberbullying detection system would involve a comprehensive assessment of its performance using standard evaluation metrics and rigorous testing procedures. The results of these evaluations would provide valuable insights into the system's effectiveness, limitations, and potential for real-world application, contributing to the ongoing effort to create safer online environments

VI. CONCLUSION

this project's exploration of a Neutrosophic Logic-enhanced MLP model for cyberbullying detection represents a crucial step towards building more resilient and context-aware automated systems. Moving beyond traditional binary logic, the integration of Neutrosophic principles allows for a more nuanced representation of the inherent uncertainties in online communication. This shift in approach not only promises improved accuracy in identifying harmful content but also offers the potential for greater explainability in model decision-making. Ultimately, this work contributes to the ongoing effort to develop intelligent tools capable of fostering safer and more inclusive digital environments, acknowledging the complexities of language and the ever-evolving nature of online interaction.

VII. REFERENCE

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/s41300-023-001860.

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. 54...

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.

10. F. Smarandache, ‘‘Neutrosophic logic—A generalization of the intuitionistic fuzzy logic,’’ *SSRN Electron. J.*, vol. 4, p. 396, Jan. 2016, doi: 10.2139/ssrn.2721587.

11. S. Das, B. K. Roy, M. B. Kar, S. Kar, and D. Pamuč ar, ‘‘Neutrosophic fuzzy set and its application in decision making,’’ *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 11, pp. 5...

19. Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. EMNLP.

20. Mikolov, T., et al. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.