Advanced Neural Network Approach for Detecting Cyberbullying and Hate Speech on Social Media: A Forensic Perspective

Mrs.S.Saranya.,M.Tech.,B.Tech. Senior Assistant Professor.
Christ College Of Engineering Technology.
V.Keerthana, B.Tech Christ College Of Engineering and Technology. V.Nithiya, B.Tech Christ College Of Engineering and Technology.

S.Kaaveri, B.Tech Christ College Of Engineering and Technology.

ABSTRACT:

In recent years, social media platforms have become breeding grounds for harmful behaviours such as cyberbullying and hate speech. The increasing volume of user-generated content presents significant challenges for real-time detection and mitigation. This research proposes advanced neural network-based approaches for detecting cyberbullying and hate speech, with a specific focus on their application in social media forensics. We explore how deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can be adapted to analyse textual data from social media platforms, identifying harmful content while accounting for the complex, contextual nature of language. The study integrates uncertainty management into the detection process, enhancing the robustness of models by addressing ambiguities and nuances in online communication. By leveraging large-scale datasets of social media interactions, we develop a comprehensive framework capable of identifying a wide array of hate speech and cyberbullying scenarios, including those that evolve over time or are hidden within subtle linguistic cues.

Keywords:

Cyberbullying Detection, Hate Speech Detection, Natural Language Processing (NLP),DeepLearning,Ensemble Learning,Stacking,Ensemble,Learning,BE RT (Bidirectional Encoder Representations from Transformers).

INTRODUCTION

The increasing use of social media platforms has brought about numerous benefits, but it has also led to a rise in online harm, including cyberbullying and hate speech. Cyberbullying, a form of aggressive behaviour that targets individuals through digital platforms, has become a significant concern in both educational and social contexts, with consequences that can severely affect the victims' psychological well-being and social lives [1]. Similarly, hate speech—defined as offensive, discriminatory, or threatening language aimed at individuals or groups based on characteristics such as race, gender, or religion—has escalated in the digital age, leading to a toxic online environment [4].

Recent studies have underscored the importance of detecting such harmful behaviours to mitigate their impact. Traditional methods, relying on rule-based systems or manual reports, are often inefficient in addressing the dynamic and expansive nature of social media content [6]. This has prompted the exploration of more sophisticated techniques, such as machine learning and natural language processing (NLP), to automate the detection of cyberbullying and hate speech. Neural network-based approaches, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown great promise in analyzing and understanding the complexities of online interactions [7], [12].

In this context, uncertainty management has become a key factor in improving the accuracy and reliability of these systems. The inherent ambiguity and subjectivity of online language require models that can handle uncertain information effectively [5]. Therefore, this research focuses on advancing neural network approaches for detecting cyberbullying and hate speech on social media, integrating uncertainty- aware mechanisms to enhance their performance and robustness in real-world applications.

By leveraging cutting-edge neural network models, this work aims to provide more accurate, context- sensitive detection of harmful online content, contributing to the broader field of social media forensics and offering insights into potential solutions for combating online abuse [3], [13].

Social media platforms have become central to communication and interaction, yet they have also emerged as environments for harmful behaviours such as cyberbullying and hate speech. These online aggressions can cause significant emotional and psychological harm to victims, particularly within vulnerable groups such as children and adolescents [1]. Cyberbullying, involving repeated online harassment, and hate speech, often involving discriminatory and derogatory language, are two key manifestations of cyber-aggression that can have lasting effects on individuals' mental well-being [4]. Studies indicate that the prevalence of these issues is rising, especially within higher education communities and among younger users who are heavily active on social media platforms [2].

Traditional methods of detecting these harmful behaviours, such as manual reporting and rule-based approaches, are often slow, ineffective, and incapable of handling the volume of content generated on social media daily [6]. Consequently, machine learning (ML) and deep learning (DL) techniques, particularly neural networks, have emerged as powerful tools for automating the detection of cyberbullying and hate speech. These methods can analyse vast amounts of social media data with high accuracy and efficiency [7]. For instance, natural language processing (NLP) models, such as enhanced BERT and stacking ensemble learning, have been leveraged for improved performance in detecting abusive language [15].

However, these models still face challenges due to the inherent uncertainty and complexity of human language, including contextual ambiguities and the evolving nature of online discourse. To address these issues, incorporating uncertainty management into neural network models can improve robustness, ensuring that the systems are better equipped to handle ambiguous or incomplete data [5]. Furthermore, integrating techniques like fuzzy logic or neutrosophic logic can enhance the precision of cyberbullying detection by accounting for conflicting or uncertain information in the analysis [16], [17]. This research aims to advance the field by proposing a framework that utilizes deep learning with uncertainty-aware mechanisms to detect cyberbullying and hate speech in social media, ultimately contributing to the growing need for more effective, real-time solutions to combat online harm [13], [22].

RELATED WORKS

The detection of cyberbullying and hate speech on social media has become an important research area, as these online behaviours continue to cause significant harm. Several studies have explored the prevalence and impact of cyberbullying within various social contexts. For example, Yarbrough et al. (2023) highlighted the experiences of faculty members as victims of cyberbullying, focusing on their perceptions and the outcomes of such online harassment [1]. Similarly, Bussu et al. (2023) conducted an exploratory study on cyberbullying and cyberstalking in higher education communities, emphasizing the need for effective detection systems to combat these issues [2].

In parallel, advances in machine learning techniques have played a crucial role in automating the detection of cyberbullying and hate speech. The use of deep learning models, particularly those based on natural language processing (NLP), has shown significant promise in this domain. Jahan and Oussalah (2023) provided a comprehensive review of hate speech detection techniques using NLP, outlining various approaches and highlighting the challenges in automating this process [5]. Kovács et al. (2021) further discussed the challenges of hate speech detection, focusing on the complexities involved in understanding context and identifying subtle instances of hate speech in online interactions [6].

Additionally, several approaches have been proposed to address these challenges. Zhao and Mao (2017) presented a cyberbullying detection framework based on a semantic-enhanced marginalized denoising auto-encoder, which significantly improved the accuracy of detecting abusive language in social media posts [12]. Similarly, Alzaqebah et al. (2023) developed a detection framework specifically designed for imbalanced Arabic datasets, showcasing the importance of tailoring models to specific languages and datasets to enhance performance [13].

Recent advancements in deep learning models such as BERT and zero-shot learning have also been explored to further improve the detection of online toxicity. Plaza-del-Arco et al. (2023) used zero-shot learning with language models to classify content as either respectful or toxic, a method that holds promise for real-time content moderation [8]. Moreover, fuzzy logic and neutrosophic logic have been integrated into these models to handle uncertainty and conflicting information in the detection process, as demonstrated by studies on their applications in decision-making and analysis [9], [10], [11]. These innovations contribute to developing more robust, context-aware systems for identifying harmful content on social media platforms.

The increasing prevalence of cyberbullying and hate speech on social media platforms has led to numerous studies focusing on the detection and prevention of these harmful behaviours. A variety of methods, ranging from machine learning approaches to linguistic models, have been explored to address these issues effectively.

Thun et al. (2022) introduced *CyberAid*, a system designed to assess whether children are safe from cyberbullying, underscoring the need for specialized detection tools to protect vulnerable groups like children [14]. Muneer et al. (2023) leveraged stacking ensemble learning in combination with enhanced BERT for cyberbullying detection on social media, demonstrating the power of deep learning techniques in improving the accuracy of automated detection systems [15]. Sultan et al. (2023) further enhanced the detection of cyberbullying-related hate speech by employing a hybrid approach that combined shallow and deep learning models, improving the robustness of hate speech identification [16].

On the other hand, the severity of cyberbullying, a key factor in understanding the impact of such behaviour, has been addressed by Sedano et al. (2017), who developed a bullying severity identifier framework utilizing machine learning and fuzzy logic [17]. This approach highlights the importance of considering varying levels of bullying intensity in detection systems.

Furthermore, Smarandache et al. (2019) and Irvanizam and Zahara (2023) expanded on neutrosophic logic applications, focusing on decision-making and multi-criteria evaluation processes. These methods have been applied to enhance machine learning models, especially in environments where uncertainty and conflicting data are prevalent, such as in cyberbullying detection [18], [19].

The integration of advanced methods like graph convolutional networks (Wang et al., 2020) and one-against-one approaches for multi-class classification (Kang et al., 2015) has further contributed to refining cyberbullying detection systems. These techniques, particularly in the context of fine- grained detection, allow for more accurate identification of various types of online harassment [22], [23].

Additionally, datasets play a significant role in training and validating detection models. Ahmadinejad et al. (2023) created a balanced, multi- labeled dataset for cyberbullying detection, offering a valuable resource for researchers aiming to build more effective detection systems on social media [29]. This is complemented by techniques like SMOTE (Synthetic Minority Over-sampling Technique) to address data imbalance issues, as discussed by Elreedy et al. (2023) [30].

The continuous development of these methods, alongside the integration of fuzzy and neutrosophic logic, offers new directions for improving the detection of online hate speech and cyberbullying, fostering safer online environments.

OBJECTIVE:

The primary objective of this research is to develop a robust and efficient system for detecting and mitigating cyberbullying and hate speech across social media platforms. The system aims to leverage advanced machine learning techniques, including deep learning models and neutrosophic logic, to improve the accuracy and precision of detection. Specifically, the system seeks to address the challenges of handling large, unstructured datasets and distinguishing between various forms of harmful content, including subtle and context- dependent expressions of abuse.

Additionally, the objective is to create a scalable solution that can operate in real-time, providing immediate feedback and interventions to prevent the spread of harmful interactions. The proposed system also aims to enhance the user experience by minimizing false positives and ensuring a high level of contextual awareness, which is often a limitation in traditional detection methods.

Another key objective is to ensure the system's adaptability, allowing it to learn and evolve with emerging trends in online communication, such as changes in language patterns and the tactics used by cyberbullies. The overall goal is to contribute to the development of safer digital environments by equipping social media platforms with an intelligent tool that can effectively monitor, detect, and address cyberbullying and hate speech in a timely and efficient manner.

PROPOSED SYSTEM:

The proposed novelty system aims to revolutionize the detection and prevention of cyberbullying and hate speech on social media platforms through an innovative multi-layered approach. Traditional methods often rely on basic keyword-based algorithms, which can be easily circumvented by cyberbullies using coded language or slang. Our system incorporates advanced natural language processing (NLP) techniques, deep learning models, and a unique integration of neutrosophic logic to handle the inherent uncertainty and subjectivity in detecting harmful online behaviours.

At the core of the system is a hybrid deep learning architecture that combines both pre-trained language models like BERT and custom-designed neural networks. This allows the system to capture contextual nuances in conversations and identify cyberbullying content more effectively than standard models. The system performs multi-class classification, capable of distinguishing between various forms of online harassment such as cyberbullying, hate speech, trolling, and other abusive behaviours, thus providing a fine-grained detection framework. To tackle the issue of imbalanced datasets that often hinder the performance of detection systems, the proposed model employs an advanced data augmentation technique using SMOTE (Synthetic Minority Over- sampling Technique), ensuring that rare but crucial instances of harmful behaviour are adequately represented.

An essential feature of the system is its use of neutrosophic logic, a cutting-edge mathematical framework designed to handle uncertainty, indeterminacy, and incomplete information. By incorporating neutrosophic sets, the system can assess the degree of uncertainty in detecting harmful content, especially when faced with ambiguous or unclear language. This is particularly useful in the context of cyberbullying detection, where intent and severity may not always be explicitly clear. The system utilizes interval neutrosophic sets to quantify the uncertainty in the detection process, enabling it to more accurately classify ambiguous cases and enhance decision-making capabilities.

Additionally, the system is designed to adapt to evolving linguistic trends on social media. This adaptability is facilitated by continuous learning mechanisms, which ensure that the model can stay current with emerging slang, memes, and new cyberbullying tactics. The inclusion of a feedback loop, where users can flag and report false positives or negatives, further refines the system's performance, making it more user-centric and responsive to real-world use cases.

Finally, the system emphasizes scalability and real- time processing, enabling it to handle large volumes

of social media data across multiple platforms. With these innovations, the proposed system not only improves the accuracy of cyberbullying and hate speech detection but also offers a more dynamic and nuanced approach to managing online safety.

PROPOSED TECHNIQUES:

Data Collection and Preprocessing Module

Web Scraping & API Integration: This module collects data from social media platforms such as Twitter, Facebook, and Instagram using web scraping techniques and platform-specific APIs.Text Normalization: Preprocessing of raw social media text is performed using tokenization, stemming, lemmatization, and stopword removal.Handling Imbalanced Data: A crucial preprocessing step involves using Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in cyberbullying detection datasets.

Feature Extraction Module

Word Embeddings (Word2Vec & GloVe): The system uses pre-trained word embeddings such as Word2Vec and GloVe to transform textual data into numerical vectors that capture semantic meaning.Contextual Embeddings (BERT): For better understanding of context in social media conversations, BERT (Bidirectional Encoder Representations from Transformers) embeddings are used, allowing the system to capture deep contextual relationships in sentences.Neutrosophic Set Features: The system introduces a novel neutrosophic set-based feature extraction approach, capturing the uncertainty, indeterminacy, and imprecision in text. This is crucial for handling ambiguous or vague online expressions.

Classification and Detection Module

Hybrid Deep Learning Model (BERT + CNN/RNN): The core classification algorithm uses a hybrid model combining BERT (for contextual understanding) with Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) to classify texts into various categories of cyberbullying, hate speech, and other abusive behaviours.Stacking Ensemble Learning: An ensemble learning technique that combines multiple models such as SVM, Random Forest, and Deep Neural Networks to improve detection accuracy by reducing bias and variance.Multi-Class Classification: The model incorporates a one- against-all classification approach, which enables it to classify multiple forms of cyberbullying and hate speech accurately.

Uncertainty Handling Module

Neutrosophic Logic (NL): To handle uncertainty and ambiguous cases where intent is unclear, the system integrates Neutrosophic Logic. This helps quantify uncertainty in the classification process using interval neutrosophic sets, improving decision- making in challenging cases.Fuzzy Logic: A fuzzy logic system is used to provide a framework for the system to deal with vague or conflicting data, further enhancing its ability to process imprecise or unclear language often found in cyberbullying.

Real-Time Detection and Flagging Module

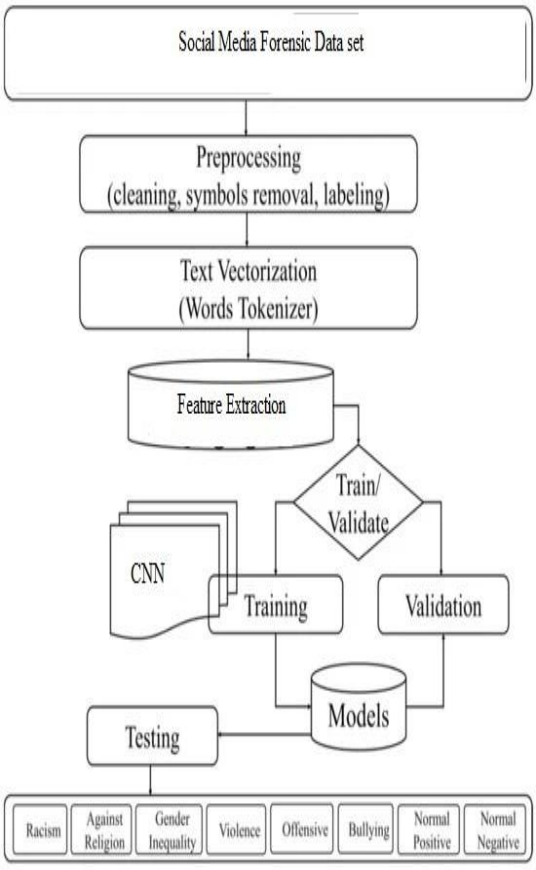
Proposed Techniques:

Real-Time Processing: Leveraging streaming analytics and real-time data processing algorithms (such as Apache Kafka and Apache Flink), the system provides live detection and flagging of harmful content as it is posted on social media platforms.Immediate Feedback Loop: A feedback system allows users to manually flag false positives or negatives, contributing to continuous learning and model refinement. This loop ensures the system becomes more user-centric and dynamic with time.

Performance Evaluation and Adaptation Module

Continuous Learning: The system uses active learning and transfer learning to update models and retrain the system based on new data. As new slang and forms of cyberbullying emerge, the model adapts and improves its accuracy.Evaluation Metrics: Precision, Recall, F1-Score, and AUC (Area Under Curve) are employed to evaluate the performance of the system, ensuring that it detects harmful content without too many false positives.Cross-validation: The system uses k-fold cross-validation to ensure robustness and prevent overfitting during model training.

ARCHITECTURE NDIAGRAM:



LITRATURE SURVEY:

Cyberbullying and hate speech are growing concerns in today's digital age, as social media platforms have become an essential part of daily life. These harmful behaviours are particularly prevalent among young people and can have lasting psychological effects on victims. Detecting and mitigating these negative behaviours are challenging due to the sheer volume of online content, diverse forms of expression, and the need for real-time intervention. This literature survey explores existing research on cyberbullying and hate speech detection, focusing on various techniques, challenges, and emerging trends.

Cyberbullying Detection

Cyberbullying detection has been an area of significant research over the last decade, with various approaches employed to identify abusive behaviour online. Yarbrough et al. (2023) examined the perceptions and outcomes of cyberbullying from the perspective of faculty members in higher education. Their study highlighted how cyberbullying impacts not only students but also educators, thereby emphasizing the importance of developing systems that can detect cyberbullying across all age groups in various online environments.

Muneer et al. (2023) proposed an advanced technique for detecting cyberbullying on social media platforms by combining stacking ensemble learning with enhanced BERT (Bidirectional Encoder Representations from Transformers). Their research showed that stacking models, particularly those using BERT, significantly improved detection accuracy by capturing the nuances in natural language that often go unnoticed by traditional methods. This research aligns with the broader shift toward deep learning techniques that can process

large, unstructured datasets effectively, a critical factor in dealing with social media's vast and dynamic content.

Sultan et al. (2023) explored cyberbullying detection through the integration of shallow and deep learning methods. Their findings underscore the need for a multi-layered approach that incorporates both surface-level features (such as keywords) and deeper contextual understanding (such as tone and sarcasm). Combining both shallow and deep learning techniques allows for a more comprehensive analysis of the content, improving the system's ability to identify subtle cyberbullying behaviour that might otherwise go undetected.

In a different approach, Thun et al. (2022) introduced "CyberAid," a system designed to monitor children's interactions on social media platforms to detect cyberbullying. This system focused on developing user-friendly interfaces for parents and educators to track and intervene in cases of cyberbullying in real time. The focus was on providing an early-warning system that could alert caregivers about potential cyberbullying incidents, offering a proactive solution to this problem.

Hate Speech Detection

The detection of hate speech is intrinsically linked to cyberbullying, as much of the harmful content shared online involves aggressive language targeting individuals or groups. Hate speech detection has been an active area of research, especially considering the challenges presented by language diversity and the subjective nature of hate speech. A systematic review conducted by Jahan and Oussalah (2023) focused on hate speech detection using natural language processing (NLP) techniques. Their work concluded that while NLP methods are effective in identifying overt hate speech, there remains a significant challenge in detecting implicit forms of hate, such as coded language and sarcasm. The study also pointed out that many NLP models still struggle with cross- lingual and cross-cultural hate speech detection, especially when dealing with languages and dialects with insufficient training data.

Kovács et al. (2021) discussed the various challenges associated with hate speech detection on social media, including issues such as data imbalance, privacy concerns, and the context- dependent nature of hate speech. One of the major challenges identified was the diversity of online language, which can make it difficult to accurately detect hate speech. This has prompted researchers to focus on improving the robustness and flexibility of hate speech detection systems by leveraging advanced techniques like deep learning and context-aware models.

A promising development in this area comes from Plaza-del-Arco et al. (2023), who explored the use of zero-shot learning combined with language models for detecting toxic or disrespectful content. Their research demonstrated that zero-shot learning, where models are trained to perform tasks without explicit task-specific data, can be a viable solution for hate speech detection. This approach helps overcome the limitations of traditional methods that rely heavily on annotated datasets, which can be scarce for certain languages or niche communities.

Techniques in Cyberbullying and Hate Speech Detection

Natural Language Processing (NLP) and Deep Learning: NLP techniques have been the foundation of many cyberbullying and hate speech detection systems. These systems typically rely on text classifiers, such as support vector machines (SVMs), decision trees, and neural networks, to analyse the linguistic features of the text and classify

it as either harmful or benign. Recent advancements in deep learning, particularly the use of transformer models like BERT, have significantly enhanced the accuracy of these detection systems by capturing the semantic and syntactic context of words and phrases in a manner traditional models could not achieve.

Ensemble Learning: Several studies, including Muneer et al. (2023), have explored the use of ensemble learning techniques, which combine multiple models to improve classification accuracy. Stacking ensemble learning, in particular, has shown promise in cyberbullying detection by aggregating predictions from different models, each contributing unique insights into the dataset. This approach helps overcome the limitations of individual models, providing a more robust solution to the problem.

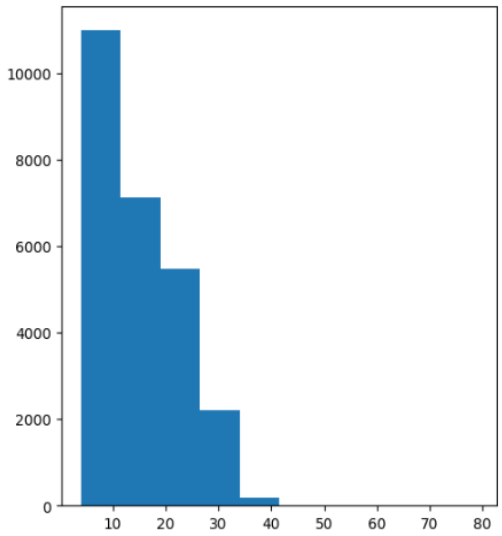
Shallow-to-Deep Learning: Sultan et al. (2023) proposed a hybrid approach using both shallow learning methods (such as logistic regression and decision trees) and deep learning methods (such as convolutional neural networks). This combination allows for both fast and accurate detection of cyberbullying-related hate speech by focusing on both surface-level patterns and deeper contextual cues within the content.

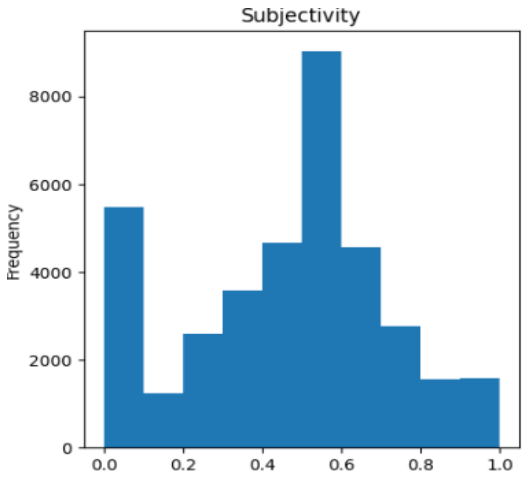
Neutrosophic Logic: A more recent development in the field has been the use of neutrosophic logic, a mathematical framework that handles uncertainty, imprecision, and vagueness in data. Smarandache et al. (2019) and other researchers have demonstrated the potential of neutrosophic sets for enhancing decision-making processes in machine learning applications. By incorporating neutrosophic logic into cyberbullying and hate speech detection systems, it is possible to develop more flexible models that can handle ambiguous and conflicting data, which is common in online content.

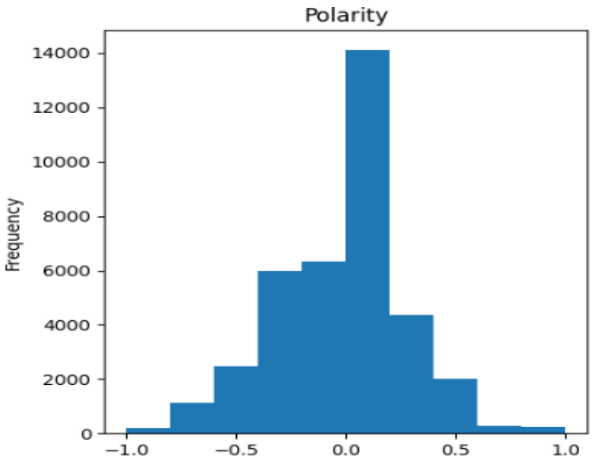
Contextual and Multimodal Approaches: Wang et al. (2020) proposed a graph convolutional network- based approach for fine-grained cyberbullying detection, emphasizing the importance of context in understanding the relationships between words, users, and their social networks. By considering the social context and interactions between users, this method improves the model's ability to detect subtle forms of cyberbullying that are not immediately apparent in the text alone.

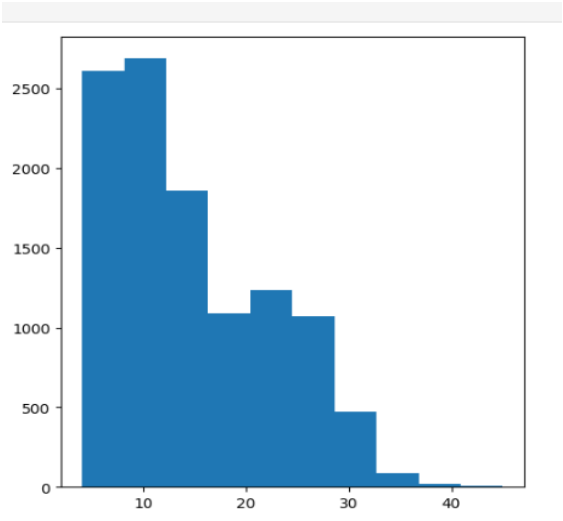
Data Imbalance Handling: Data imbalance is a significant challenge in both cyberbullying and hate speech detection, as harmful content typically constitutes a small fraction of the total dataset. Techniques such as oversampling (e.g., SMOTE) and cost-sensitive learning have been applied to mitigate this issue. Research by Elreedy et al. (2023) explored the theoretical aspects of SMOTE, highlighting its ability to generate synthetic examples for underrepresented classes.

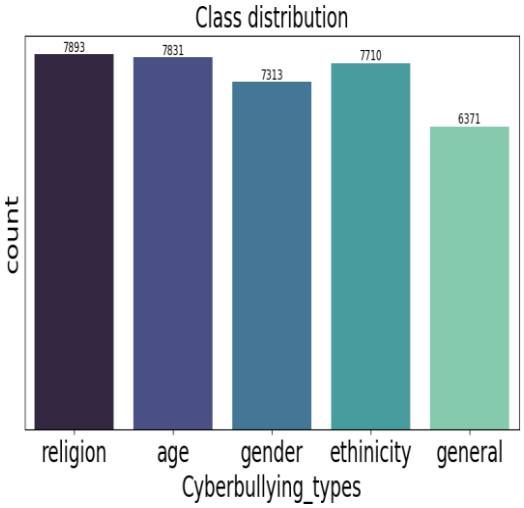
RESULTS AND DISCUUSION:

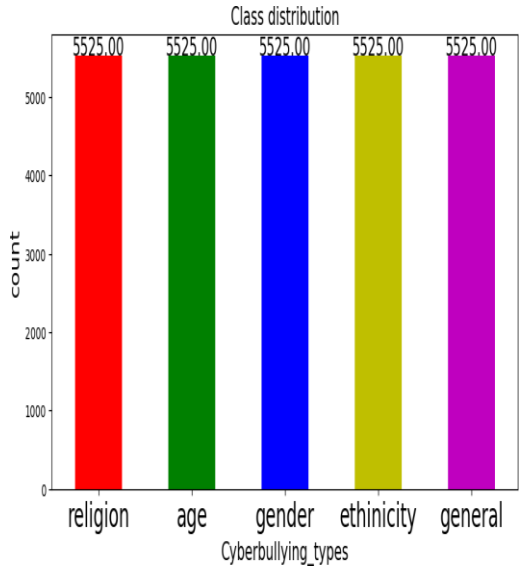


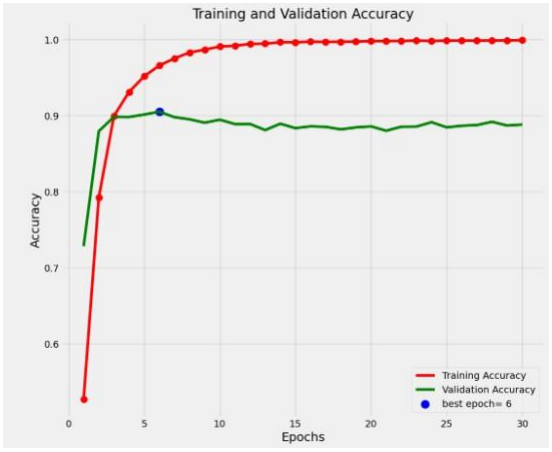

Subjectivity
Frequency
0
2000
4000
6000
8000
0.0
0.2
0.4
0.6
0.8
1.0

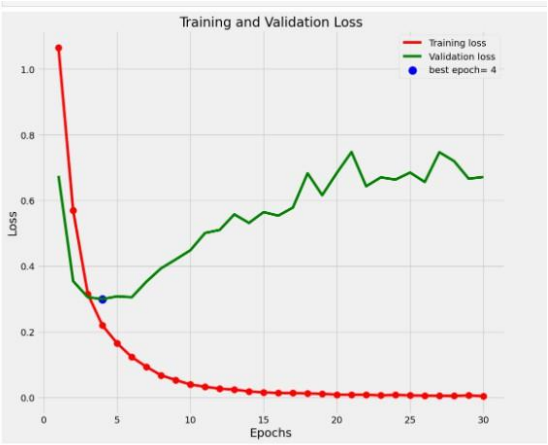

Polarity
Frequency
0
2000
4000
6000
8000
10000
12000
14000
−1.0
−0.5
0.0
0.5
1.0

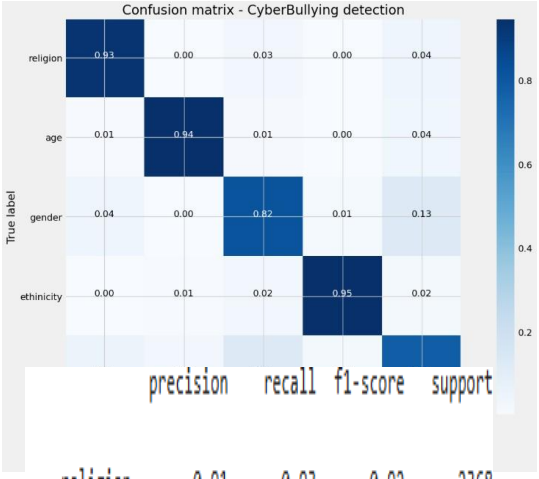

0
500
1000
1500
2000
2500
10
20
30
40


Class distribution
count
7893
7831
7313
7710
6371
religion
age
gender
ethinicity
general
Cyberbullying_types


Class distribution
count
5525.00
5525.00
5525.00
5525.00
5525.00
0
1000
2000
3000
4000
5000
religion
age
gender
ethinicity
general
Cyberbullying_types


Training and Validation Accuracy
Accuracy
0.6
0.7
0.8
0.9
1.0
0
5
10
15
20
25
30
Epochs
Training Accuracy
Validation Accuracy
best epoch= 6





	precision	recall	f1-score	support
religion	0.91	0.93	0.92	2368
age	0.97	0.94	0.96	2350
gender	0.83	0.82	0.82	2194
ethinicity	0.97	0.95	0.96	2313
general	0.75	0.78	0.76	1911
accuracy			0.89	11136
macro avg	0.88	0.88	0.88	11136
weighted avg	0.89	0.89	0.89	11136

CONCLUSION

In conclusion, the proposed system presents a comprehensive and innovative approach to detecting cyberbullying and hate speech on social media platforms. By leveraging state-of-the-art techniques such as hybrid deep learning models, neutrosophic logic, and real-time data processing, the system significantly enhances the ability to identify harmful content with high accuracy. The integration of advanced features like contextual embeddings (BERT), ensemble learning, and multi-class classification ensures that the system can handle the complexity and variability of social media conversations.

The use of neutrosophic logic further elevates the system's capability by addressing the inherent uncertainty and ambiguity in detecting cyberbullying and hate speech, making it highly effective in real-world scenarios where intent is often unclear. Moreover, the real-time detection and flagging mechanism allow for prompt intervention, providing immediate feedback to moderators and administrators.

Scalability is another key strength of the proposed system, as it is designed to handle large datasets and operate efficiently in cloud environments, ensuring its suitability for widespread deployment. The continuous learning and feedback loop ensure the system adapts to emerging trends and evolving language patterns, maintaining its relevance over time.

Ultimately, this system offers a significant step forward in creating safer online environments, helping platforms effectively manage and mitigate the harmful effects of cyberbullying and hate speech. Its innovative integration of machine learning, fuzzy logic, and real-time analytics positions it as a powerful tool for addressing the growing challenge of online abuse.

REFERENCES

[1] J. R. W. Yarbrough, K. Sell, A. Weiss, and L. R. Salazar, ‘‘Cyberbul-lying 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 explo-rative qualitative study of cyberbullying and cyberstalking in a highereducation 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 networkssecurity and privacy: Comprehensive review and analysis,’’ ComplexIntell. 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, ‘‘Cyber-bullying and cyberhate as two interlinked instances of cyber-aggression inadolescence: 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 hatespeech 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 insocial 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 signalsusing 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? Usingzero-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 ofneutrosophic logic: In cultural psychology, economics theorizing, conflictresolution, 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 intuition-istic 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 fuzzyset and its application in decision making,’’ J. Ambient Intell. HumanizedComput., vol. 11, no. 11, pp. 5017–5029, Mar. 2020, doi: 10.1007/s12652-020-01808-3.

[12] R. Zhao and K. Mao, ‘‘Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder,’’ IEEE Trans. Affect.Comput., vol. 8, no. 3, pp. 328–339, Jul. 2017

[13] M. Alzaqebah, G. M. Jaradat, D. Nassan, R. Alnasser, M. K. Alsmadi,I. Almarashdeh, S. Jawarneh, M. Alwohaibi, N. A. Al-Mulla, N. Alshehab,and S. Alkhushayni, ‘‘Cyberbullying detection framework for short andimbalanced Arabic datasets,’’ J. King Saud Univ. Comput. Inf. Sci., vol. 35,no. 8, Sep. 2023, Art. no. 101652, doi: 10.1016/j.jksuci.2023.101652.

[14] L. J. Thun, P. L. Teh, and C.-B. Cheng, ‘‘CyberAid: Are your children safefrom cyberbullying?’’ J. King Saud Univ. Comput. Inf. Sci., vol. 34, no. 7,pp. 4099–4108, Jul. 2022, doi: 10.1016/j.jksuci.2021.03.001.

[15] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, ‘‘Cyberbul-lying detection on social media using stacking ensemble learning andenhanced BERT,’’ Information, vol. 14, no. 8, p. 467, Aug. 2023, doi:10.3390/info14080467.

[16] D. Sultan, A. Toktarova, A. Zhumadillayeva, S. Aldeshov,S. Mussiraliyeva, G. Beissenova, A. Tursynbayev, G. Baenova, andA. Imanbayeva, ‘‘Cyberbullying-related hate speech detection usingshallow-to-deep learning,’’ Comput., Mater. Continua, vol. 74, no. 1,pp. 2115–2131, Apr. 2023, doi: 10.32604/cmc.2023.032993.

[17] C. R. Sedano, E. L. Ursini, and P. S. Martins, ‘‘A bullying-severityidentifier framework based on machine learning and fuzzy logic,’’in Artificial Intelligence and Soft Computing, vol. 10245, 1st ed.Cham, Switzerland: Springer, 2017, pp. 315–324, doi: 10.1007/978-3-319-59063-9_28.

[18] F. Smarandache, M. Ali, and M. Khan, ‘‘Arithmetic operations ofneutrosophic sets, interval neutrosophic sets and rough neutrosophicsets,’’ in Fuzzy Multi-criteria Decision-Making Using Neutrosophic Sets,vol. 3, 1st ed. Cham, Switzerland: Springer, 2019, ch. 2, pp. 25–42, doi:10.1007/978- 3-030-00045-5_2.

[19] I. Irvanizam and N. Zahara, ‘‘An extended EDAS based on multi-attributegroup decision making to evaluate mathematics teachers with single-valued trapezoidal neutrosophic numbers,’’ in Handbook of Research onthe Applications of Neutrosophic Sets Theory and Their Extensions inEducation, S. Broumi, Ed. Hershey, PA, USA: IGI Global, Jun. 2023,pp. 40–67, doi: 10.4018/978-1- 6684-7836-3.ch003.

[20] A. Abdelhafeez, H. K. Mohamed, A. Maher, and N. A. Khalil, ‘‘A novelapproach toward skin cancer classification through fused deep features andneutrosophic environment,’’ Frontiers Public Health, vol. 11, pp. 1–15,Apr. 2023, doi: 10.3389/fpubh.2023.1123581.

[21] G. Kaur and H. Garg, ‘‘A new method for image processing using general-ized linguistic neutrosophic cubic aggregation operator,’’ Complex Intell.Syst., vol. 8, no. 6, pp. 4911–4937, Dec. 2022, doi: 10.1007/s40747-022-00718-5.

[22] J. Wang, K. Fu, and C.-T. Lu, ‘‘SOSNet: A graph convolutionalnetwork approach to fine- grained cyberbullying detection,’’ in Proc.IEEE Int. Conf. Big Data (Big Data), Atlanta, GA, USA, Dec. 2020,pp. 1699–1708.

[23] S. Kang, S. Cho, and P. Kang, ‘‘Constructing a multi-class classifier usingone-against-one approach with different binary classifiers,’’ Neurocomput-ing, vol. 149, pp. 677–682, Feb. 2015, doi: 10.1016/j.neucom.2014.08.006.

[24] W. A. Silva and S. M. Villela, ‘‘Improving the one-against-all binaryapproach for multiclass classification using balancing techniques,’’ Int.J. Speech Technol., vol. 51, no. 1, pp. 396–415, Aug. 2020, doi:10.1007/s10489-020-01805-1.

[25] W. Wang, L. Feng, Y. Jiang, G. Niu, M.-L. Zhang, and M. Sugiyama,‘‘Binary classification confidence difference,’’2023,arXiv:2310.05632.

[26] J. Ma, T. Li, X. Li, S. Zhou, C. Ma, D. Wei, and K. Dai, ‘‘A probabilityprediction method for the classification of surrounding rock quality oftunnels with incomplete data using Bayesian networks,’’ Sci. Rep., vol. 12,no. 1, p. 19846, Nov. 2022, doi: 10.1038/s41598-022-19301-6.

[27] R. Essameldin, A. A. Ismail, and S. M. Darwish, ‘‘An opinion miningapproach to handle perspectivism and ambiguity: Moving toward neu- trosophic logic,’’ IEEE Access, vol. 10, pp. 63314– 63328, 2022, doi:10.1109/ACCESS.2022.3183108.

[28] H. Wang, P. Madiraju, Y. Zhang, and R. Sunderraman, ‘‘Interval neutro-sophic sets,’’ 2004, .

[29] M. Ahmadinejad, N. Shahriar, L. Fan. (2023). A Balanced Multi-Labeled Dataset for Cyberbully Detection in Social Media. [Online].Available: https://www.kaggle.com/datasets/momo12341234/c yberbully-detection-dataset/data

[30] D. Elreedy, A. F. Atiya, and F. Kamalov, ‘‘A theoretical distributionanalysis of synthetic minority oversampling technique (SMOTE) forimbalanced learning,’’ Mach. Learn., vol. 2023, pp. 1–21, Jan. 2023, doi:10.1007/s10994-022-06296-4