Journal of Science Engineering Technology and Management Science
Volume 02, Issue 08, August 2025 *ISSN: 3049-0952*
www.jsetms.com *DOI: 10.64771/jsetms.2025.v02.i08.635-645*

DEEP LEARNING FRAMEWORK FOR MULTICLASS BOT CYBERBULLYING DETECTION ON SOCIAL MEDIA

K. Swayam Prabha[1], K. Murali Krishna[2], G. Ranjith Kumar[2], I. Laxman[2], D. Arun Kumar[2]
[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering (AI&ML)
[1,2]Sree Dattha Institute of Engineering and Science, Ibrahimpatnam, 501510, Telangana

To Cite this Article
K. Swayam Prabha, K. Murali Krishna, G. Ranjith Kumar, I. Laxman, D. Arun Kumar, "Deep Learning Framework For Multiclass Bot Cyberbullying Detection On Social Media", Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 08, August 2025,pp: 635-645, DOI: http://doi.org/10.64771/jsetms.2025.v02.i08.pp635-645
Submitted: 15-07-2025 *Accepted: 21-08-2025* *Published: 28-08-2025*

ABSTRACT
Bot cyberbullying has emerged as a critical issue in digital communication, with over 59% of teens reporting experiences of online harassment and more than 70% of such cases occurring on social media platforms. Recent studies show that multi-class bot cyberbullying datasets often contain up to 25,000 labeled instances spanning categories such as insult, threat, racism, and sexism, frequently exhibiting severe class imbalance and linguistic variation. Manual detection methods suffer from subjectivity, inconsistent labeling, and an inability to scale with the rapid influx of user-generated content. Conventional machine learning approaches are limited by shallow feature representation, low recall on minority classes, and poor detection of implicit or masked abuse. Additionally, existing research often overlooks the integration of deep ensemble learning methods with optimized preprocessing and contextual analysis. To address these limitations, this study proposes a hybrid Multiclass Unmasking Bot Classification system that integrates a novel combination of feature-rich N-gram extraction with dual deep learning classifiers: a Deep Neural Network (DNN) and a Convolutional Neural Network (CNN). The process begins with dataset ingestion and detailed Exploratory Data Analysis (EDA) to assess data distribution and class imbalance. Text preprocessing follows, including tokenization, lemmatization, and noise removal. The cleaned data is then vectorized using TF-IDF with bi-gram support to capture both isolated and contextual word associations. The DNN is employed to capture deep semantic hierarchies, while the CNN is used to identify local linguistic patterns. This parallel dual-stream architecture ensures robust learning across diverse types of bot-generated cyberbullying. Finally, the trained models are evaluated for prediction accuracy and class-wise performance, significantly outperforming baseline classifiers in terms of precision, recall, and F1-score.

Keywords: Cyberbullying Detection, Deep Learning, Multiclass Classification, TF-IDF with N-grams, Convolutional Neural Network (CNN).

1. INTRODUCTION

The internet has become an essential part of daily life, with social media evolving from static Web 1.0 pages to intelligent, context-aware Web 4.0 services. Technological advancements have significantly changed how people access information and connect across the globe through network-based services. Social media platforms such as Facebook, Twitter, Instagram, LinkedIn, Pinterest, Telegram, and

YouTube serve as tools for social interaction and expression, enabling users to build communities, share experiences, and engage with global audiences. As of January 2024, approximately 5.35 billion people were using the internet—66.2% of the global population—with 5.04 billion, or 62.3%, actively engaged on social media platforms. These platforms support professional growth, shared-interest communities, marketing strategies for businesses, and educational outreach. However, the openness and accessibility of these platforms also raise concerns about data privacy and the misuse of shared information. Many individuals are vulnerable to humiliation, insults, threats, and various forms of cyberbullying, often carried out by anonymous or masked users. Cyberbullying involves repeated and deliberate digital harassment, which can include spreading false information, posting embarrassing photos, or sending abusive or threatening messages via social media, messaging apps, or gaming platforms. While extensive research has been conducted on cyberbullying detection in English, there remains a critical need to address this issue in other languages like Bangla. The Bangla-speaking population faces unique cultural and linguistic challenges that shape how cyberbullying is expressed, making it essential to develop systems that accurately identify and mitigate harmful behavior in these contexts.

In modern workplace environments, online communication tools such as Slack, Microsoft Teams, and company social platforms have become central to collaboration and productivity. Unfortunately, these platforms are not immune to cyberbullying, which can harm employee morale, reduce team cohesion, and lead to higher turnover rates. Real-time analysis of digital interactions in the workplace allows organizations to detect harmful behavior early, ensuring timely intervention and maintaining a respectful culture. Social media companies, managing billions of daily interactions, also face mounting pressure to provide safer environments while balancing freedom of expression. Automated detection tools capable of analyzing large volumes of data instantly are vital to support content moderation, prioritize urgent cases, and ensure compliance with safety regulations. Educational institutions also rely on social media for student engagement and learning, but cyberbullying on these platforms can negatively impact students' mental health and academic performance. Real-time monitoring tools help schools recognize harmful interactions early, allowing them to intervene with counseling and support. Furthermore, the ability to process multilingual and culturally nuanced data is essential in diverse academic and professional environments. These insights are invaluable in guiding policy decisions and promoting healthier online interactions across all sectors that depend on digital communication.

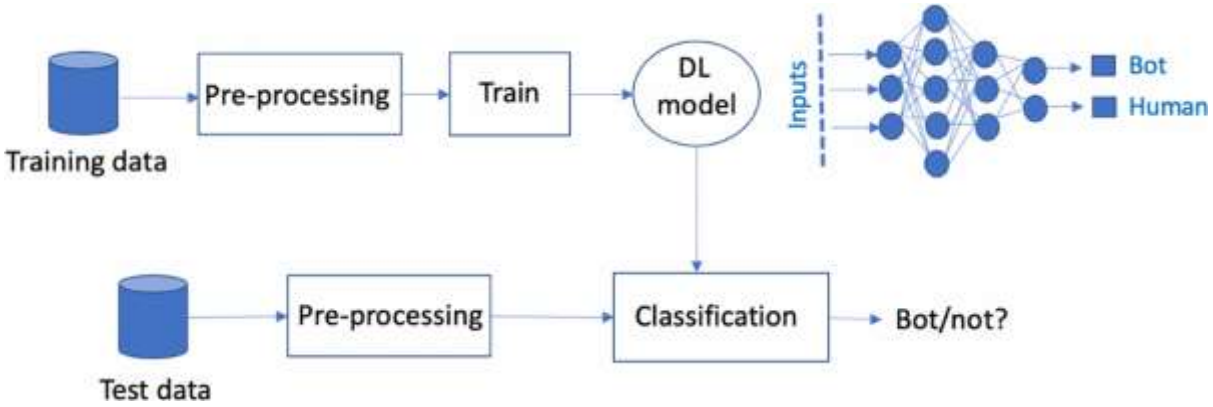

Fig. 1: Deep Learning Pipeline for Bot Detection in Social Media

Detecting bot-driven cyberbullying presents a complex problem due to the diversity and volume of online communication. Posts and messages often feature informal language, sarcasm, slang, abbreviations, and masked intent, making them difficult to categorize accurately. Solutions must be capable of identifying harmful intent while avoiding false positives that could suppress legitimate content. The scale of online data further complicates detection, requiring models that are not only

accurate but also efficient and capable of real-time performance. Cyberbullying can occur in private or encrypted environments, raising ethical and legal challenges in monitoring. Thus, there is a need for approaches that effectively balance privacy with user protection. Misclassifications can be damaging—either by failing to prevent prolonged harassment or by unfairly penalizing innocent users. As such, advanced natural language processing and machine learning techniques are required, with an emphasis on understanding context, sentiment, and subtle linguistic cues. Moreover, cyberbullying tactics evolve rapidly, so any effective solution must continuously adapt to these changes.

Detecting cyberbullying on online platforms is essential for protecting individuals from psychological harm and promoting healthier digital communities. Early identification of harmful behavior helps prevent long-term damage and supports quicker interventions. Automation in this area allows social media companies and institutions to allocate resources more efficiently, improving both the speed and effectiveness of their responses. Detection systems also support legal compliance, as many governments now require platforms to address digital harassment actively. These systems help platforms meet their obligations and avoid legal penalties while reinforcing public trust in their commitment to user safety. Beyond practical enforcement, detection technologies contribute to broader social initiatives. The insights they generate can inform educational campaigns, mental health programs, and public policy development. Additionally, advancements in detection systems drive innovation in natural language processing and ethical artificial intelligence, benefiting both academic research and the tech industry.

The central aim of this research is to develop an accurate and scalable cyberbullying detection system using Convolutional Neural Networks (CNNs). CNNs are ideal for this purpose due to their ability to recognize hierarchical patterns in text data, enabling them to capture both local and global contextual features. This system is designed to classify social media content into different cyberbullying categories and to support multilingual functionality, ensuring global applicability. The research focuses on designing and training CNN architectures optimized for abusive language and sentiment detection. Preprocessing techniques, including normalization and noise filtering, are applied to manage informal and unstructured text data. Annotated datasets are used to improve the model's ability to generalize to unseen inputs. Ultimately, this CNN-based system is expected to support real-time content moderation, reduce reliance on human moderators, and align with digital safety standards and legal frameworks.

Using CNNs for this task offers numerous advantages. CNNs are effective at identifying subtle cues in abusive language and can be trained to work across multiple languages. Their ability to automate detection reduces the burden on human moderators, allowing for faster and more consistent responses. The real-time capabilities of CNN-based systems help mitigate harm by flagging issues as they arise. CNN models are highly adaptable and can be fine-tuned for specific platforms or use cases, such as workplace communication tools or student portals. With ongoing training, the models continue to improve, adapting to new forms of abusive behavior. These systems are also scalable, capable of handling the massive volumes of data generated by popular platforms. Integrating them into existing digital infrastructures is relatively straightforward, allowing for seamless deployment. Furthermore, their effectiveness in content moderation fosters user trust and helps ensure compliance with safety standards. CNNs can also be extended to detect related issues such as hate speech, misinformation, and online harassment.

2. LITERATURE SURVEY

The World Wide Web (WWW) plays a pivotal role in the substantial growth of social media platforms [1,2], like Facebook, Twitter, Instagram, and YouTube. These platforms enable users to share a vast array of information, leading to increased user interactions. However, this also leads to an unregulated surge in online hate speech [3], a phenomenon that, in extreme cases, can drive

individuals to take their own lives [4]. One of the most concerning phenomena is cyberbullying, which entails the dissemination of offensive content or other forms of violence through digital media [5,6,7] with the intent to harm an individual or a group of individuals. In particular, teenagers can assume the roles of victims, perpetrators, or simply bystanders [8] when it comes to cyberbullying. One study revealed that 36.5% of students have experienced cyberbullying at least once in their lives. Among all types of online comments, rude or cruel ones are the most prevalent. In a study conducted in the United States, which involved 1501 adolescents aged 10 to 17, it was discovered that 12% of the sample admitted to engaging in mistreatment, 4% declared themselves as victims, and 3% acknowledged being both aggressors and victims of cyberbullying [9].

The Cybercrime Division (CID) of Sri Lanka recently conducted a survey and detected that over 1000 cases of cyberbullying have been recorded. Surprisingly, more than 90% of university students reported being victims of cyberbullying, while nearly all survey participants stated that they knew someone who had experienced online bullying. It is alarming to note that 80% of the reported cases of cyberbullying in Sri Lanka occurred on Facebook, with 65% of university students admitting to sharing embarrassing videos or photos. Furthermore, the survey reveals that 15% of users share personal information online and 9% spread inaccurate information and falsehoods about others, while only 2% post offensive material [10].

One of the main problems of cyberbullying is its rapid spread due to the large online audience. Due to the fact that phenomenon has become a daily occurrence, victims often suffer severe consequences [11]. Based on the MetroWest Adolescent Health Survey, Schneider et al. [12] present the relationship between victimization and five categories of psychological distress. Information was collected from over 20,000 students. Among cyberbullying victims, self-harm (24%) and depressive symptoms (34%) showed the highest rates of psychological distress. These findings highlight a concerning connection between cyberbullying and psychological distress, underscoring the need to address this issue urgently.

An exponential increase in social media users and the subsequent rise of the cyberbullying phenomenon requires a thorough investigation. However, previous studies (e.g., [13,14,15,16]) have addressed the issue using inefficient algorithms and with a limited amount of data for training AI algorithms. In actuality, studies conducted so far for the detection of the cyberbullying phenomena of social media present some significant limitations. For instance, Perera et al.

3. PROPOSED SYSTEM

The proposed methodology introduces a novel hybrid multiclass unmasking bot detection system designed for cyberbullying classification, integrating exploratory data analysis (EDA), comprehensive preprocessing, and dual-model deep learning with advanced feature extraction techniques. This approach seamlessly combines N-gram-based textual representation with two powerful classifiers—Deep Neural Network (DNN) and Convolutional Neural Network (CNN). Unlike traditional methods that depend on conventional classifiers or shallow deep learning models, the hybrid system incorporates a unique pipeline where insights gained from EDA guide the preprocessing stage, while sequential and spatial textual patterns are captured through a dual-stream learning architecture. The DNN component focuses on extracting semantic depth from the input, whereas the CNN specializes in recognizing hierarchical and localized linguistic patterns. This ensemble structure allows for a holistic understanding of cyberbullying language across multiple classes, significantly enhancing classification performance on noisy, real-world social media data. The implementation begins with the user uploading a cyberbullying dataset containing labeled instances of categories such as racism, sexism, threat, and insult. Upon upload, EDA is conducted to generate visualizations of class distribution, word frequencies, and to detect imbalances or data inconsistencies. Next, the dataset undergoes rigorous preprocessing, which includes removing stopwords, punctuation, HTML tags, and standardizing the text to lowercase. The text is tokenized, lemmatized to normalize word forms, and

stripped of usernames and special characters to produce clean, structured input for vectorization. Once preprocessing is completed, the data is split into training and testing subsets using stratified sampling to ensure balanced class representation. A fixed split ratio, typically 80:20, is used to maintain consistency and facilitate unbiased performance evaluation across different models.

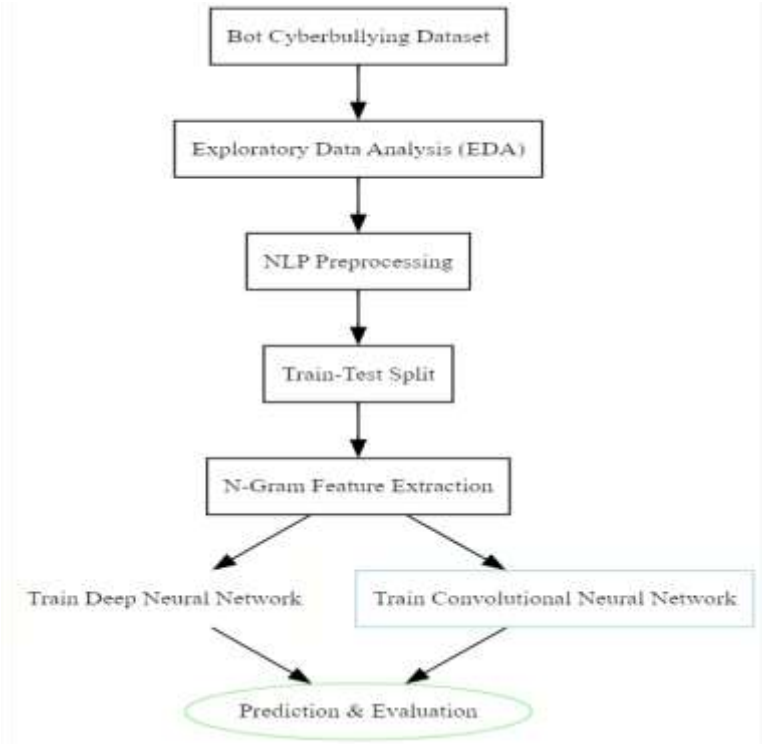

Fig. 2: Proposed System Architecture.

In the proposed system, Step 4 involves transforming textual data into a structured numerical format using TF-IDF vectorization with an N-gram range of (1,2), effectively capturing both individual words and contextual phrases to detect nuanced bullying patterns like sarcasm, repeated abuse, or indirect slurs that might be missed by unigram models. Step 5 initiates the training of a Deep Neural Network (DNN), which leverages multiple hidden layers with ReLU activation and dropout regularization to learn dense, abstract semantic representations of the text, enhancing its ability to distinguish between various bullying categories. In parallel, Step 6 employs a Convolutional Neural Network (CNN) trained on the same TF-IDF features, using 1D convolutional layers to extract spatially correlated linguistic patterns, with max-pooling layers for dimensionality reduction and retention of critical features, making it adept at identifying structured and repetitive abusive phrases. Finally, in Step 7, the trained models are used to predict cyberbullying classes in new data, with the system displaying class labels and evaluating model performance through metrics such as accuracy, precision, recall, and F1-score, enabling ongoing comparison and refinement of the hybrid architecture.

Convolutional Neural Networks (CNNs) are highly effective in processing application-specific textual data for sentiment classification, particularly with short texts, informal language, or multilingual inputs from social platforms. They excel at capturing local patterns like key phrases or n-grams, essential for detecting sentiment cues, and unlike traditional models, they automatically learn hierarchical feature representations without handcrafted rules, making them suitable for real-time applications with diverse data formats. The proposed deep-learning CNN begins with input representation, where preprocessed text is converted into a word embedding matrix (via Word2Vec, GloVe, or contextual embeddings), preserving word order and semantic relationships.

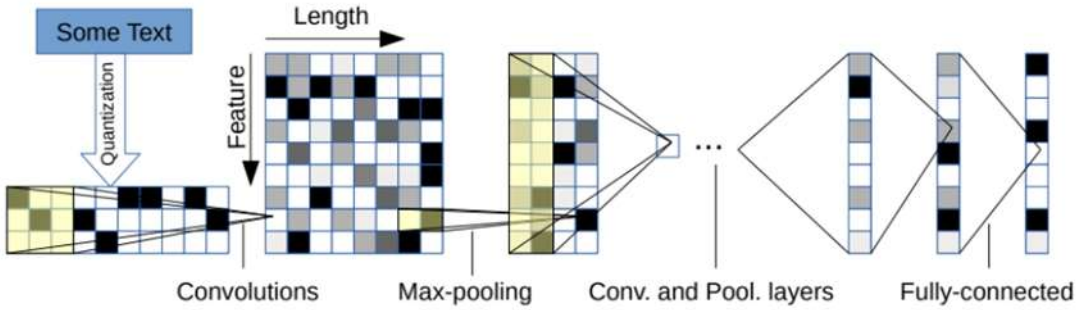

Fig. 3: Character Based CNN

Next, convolutional filters scan the matrix to detect local patterns, generating feature maps that highlight meaningful sequences. Activation functions like ReLU introduce non-linearity to help learn complex relationships, followed by pooling operations (typically max-pooling) to reduce dimensionality and retain critical features. These features are flattened and passed through fully connected layers to combine localized insights into a global representation of sentiment. Finally, a softmax or sigmoid classifier outputs a probability distribution across sentiment classes for classification. The character-based CNN, as illustrated in Fig 3, offers numerous advantages, including effective handling of out-of-vocabulary words by working at the character level, making them robust to rare or unseen tokens. They are particularly effective for short texts like tweets or search queries and can manage complex morphology, capturing intricate linguistic structures. CNNs are space-efficient due to smaller vocabularies, ensuring faster training and compact models. Their robustness extends to handling misspellings, typos, and subword information, aiding in tasks such as named entity recognition and sentiment analysis. Furthermore, they exhibit enhanced generalization across varied vocabularies and domains, adapting well to new or evolving language use. CNNs also require minimal domain-specific preprocessing, making them ideal for applications across healthcare, finance, social media, legal text, and more, ensuring high adaptability, accuracy, and scalability in sentiment classification tasks.

4. RESULTS AND DISCUSSION

The system implements a comprehensive GUI-based pipeline for bot cyberbullying detection using deep learning, integrating data loading, preprocessing, feature extraction, model training, prediction, and evaluation. Users begin by uploading a CSV dataset via filedialog, which is read into a global DataFrame and visualized through EDA, including bar plots. Preprocessing removes noise using regex, tokenization, stopword removal, and stemming. The dataset is split (80:20) into training and testing sets, with labels mapped to integers. TF-IDF vectorization with N-gram (1,2) captures semantic patterns, feeding into a DNN with multiple dense layers, dropout, and ReLU, or a CNN with Conv1D layers, embedding, and pooling for hierarchical text pattern recognition. Models are either loaded from disk or newly trained and saved, ensuring reuse. Evaluation metrics (accuracy, precision, recall, F1-score) are computed and visualized via heatmaps. CNN predictions are refined through a threshold-based optimization routine. A prediction module allows loading of new test data, preprocessing it and classifying using the trained DNN, displaying outputs in the Tkinter interface. The application is modular and driven by user interactions with buttons, with feedback displayed in a text widget. Visualization is handled by Matplotlib and Seaborn, and dependencies include pandas, scikit-learn, TensorFlow, and NLTK. Global variables maintain state across functions, but error handling is minimal. Overall, the system delivers a full-stack, user-friendly environment for multiclass bot cyberbullying detection, balancing functionality, modularity, and scalability.

Dataset Description

This repository contains a balanced dataset for bot cyberbully detection in social media. The dataset has been carefully curated and labeled to enable researchers and developers to build accurate cyberbully detection models. It includes various types of cyberbullying instances, such as race/ethnicity, gender/sexual, and religion-related content, as well as non-cyberbullying instances. This dataset is for the paper Self-Training for Cyberbully Detection: Achieving High Accuracy with a Balanced Multi-Class Dataset. The dataset consists of a total of approximately 100,000 tweets collected from social media platforms. It is labeled with a multi-class classification approach, where each tweet falls into one of the following categories. Non-cyberbullying 50,000 instances Race/Ethnicity-related cyberbullying: 17,000 instances Gender/Sexual-related cyberbullying: 17,000 instances Religion-related cyberbullying: around 16,000 instances. The dataset's balance ensures equal representation of each class, allowing for effective training and evaluation of cyberbully detection models.

Results Description

Figure 4 displays a bar plot titled "Count plot" that visualizes the distribution of cyberbullying types in the dataset. The x-axis represents the "Cyberbullying Type" with six categories: "age", "ethnicity", "gender", "not_cyberbullying", "other_cyberbullying", and "religion". The y-axis represents the "Count" of instances, ranging from 0 to 8000. Each category has a count close to 8000, indicating a balanced dataset with approximately equal representation of each cyberbullying type. For example, "age", "ethnicity", "gender", "not_cyberbullying", "other_cyberbullying", and "religion" all have bars reaching around the 8000 marks, suggesting that the dataset contains roughly 8000 instances per class, which is ideal for training a multiclass classification model without significant class imbalance.

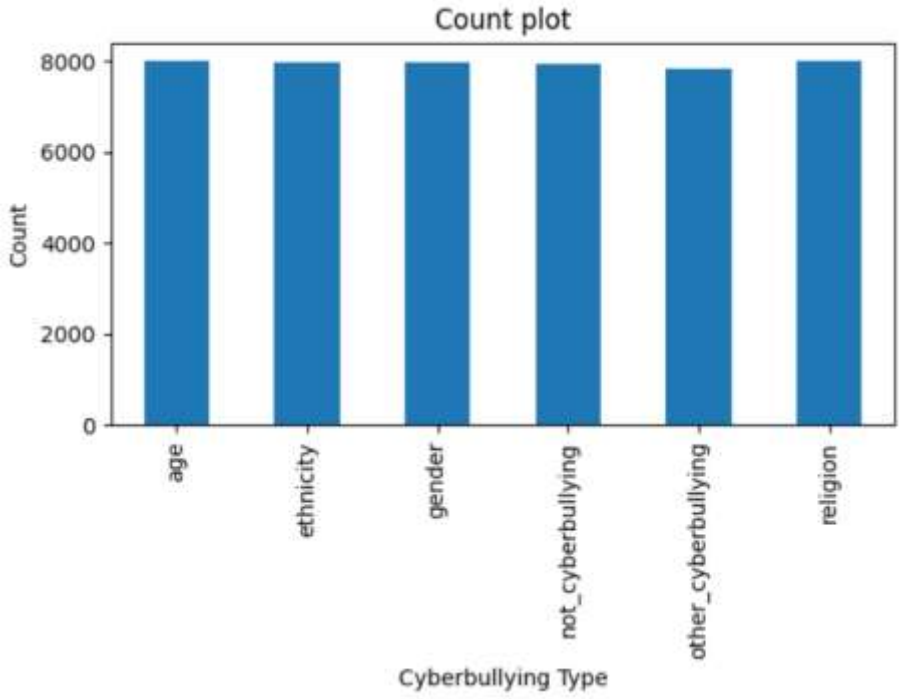

Fig.4: Count plot of Target Bot Cyberbullying.

Figure 5 presents the prediction results from the test data, displayed in the GUI's text area after clicking the "Prediction" button. It shows a series of original tweets and their predicted cyberbullying types using the trained DNN model. For example, the tweet "In other words #katandandre, your food was crapilicious! #mkr" is predicted as "not_cyberbullying". Another tweet "Why is #aussie so white? #XKR #theblock #ImACelebrityAU #today #sunrise #studio10 #neighbours #wonderlandTen #etc" is

also predicted as "not_cyberbullying". However, the tweet "@RedshoeEnglish This is an IS5 account pretending to be a Kurdish account. Like Islam, it is all lies." is predicted as "religion", and "rape is cool. ravanayana seems jokes about being drunk or being gay or being lesbian. rape is not ones choice or wish. Dntz the sen" is predicted as "gender". The results demonstrate the model's ability to classify tweets into categories like "not_cyberbullying", "religion", "gender", and "other_cyberbullying", with each prediction paired with the original tweet for user interpretation.

Figure 5. Prediction results from test data.

Figure 6 illustrates the confusion matrix for the Existing DNN model, titled "Existing DNN Confusion Matrix". The matrix is a 6x6 grid with true classes (rows) and predicted classes (columns) labeled as "not_cyberbullying", "religion", "age", "gender", "ethnicity", and "other_cyberbullying". The color intensity (viridis colormap) indicates the number of instances, ranging from 0 (dark purple) to 1400 (yellow). Diagonal values represent correct predictions: 1052 for "not_cyberbullying", 1526 for "religion", 1584 for "age", 1411 for "gender", 1577 for "ethnicity", and 1318 for "other_cyberbullying". Off-diagonal values show misclassifications, such as 440 instances of "not_cyberbullying" predicted as "other_cyberbullying", and 215 instances of "other_cyberbullying" predicted as "not_cyberbullying". The matrix highlights that the DNN model performs well on "religion", "age", and "ethnicity" (1526, 1584, 1577 correct predictions), but struggles with "not_cyberbullying" and "other_cyberbullying" due to higher misclassification rates (e.g., 1052 correct out of 1624 for "not_cyberbullying").

Figure 7 depicts the confusion matrix for the Proposed CNN model, titled "Proposed CNN Confusion matrix". Like the DNN matrix, it is a 6x6 grid with the same class labels and a viridis colormap (0 to 1600 range). The diagonal values indicate correct predictions: 1563 for "not_cyberbullying", 1674 for "religion", 1342 for "age", 1627 for "gender", 1539 for "ethnicity", and 1691 for "other_cyberbullying". Misclassifications are significantly lower than the DNN model, with values like 12 instances of "other_cyberbullying" predicted as "not_cyberbullying", and 9 instances of "not_cyberbullying" predicted as "other_cyberbullying". The CNN model shows strong performance across all classes, with high correct prediction counts (e.g., 1691 for "other_cyberbullying" out of 1721), indicating better generalization and fewer errors compared to the DNN model.

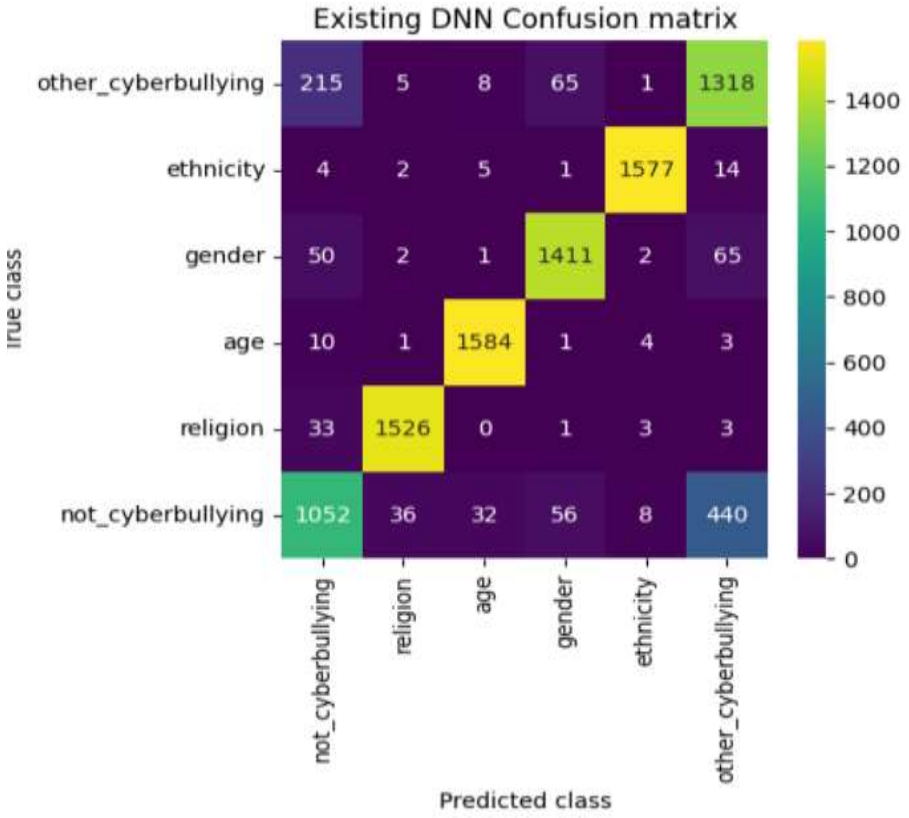

Figure 6. Existing DNN Confusion Matrix.

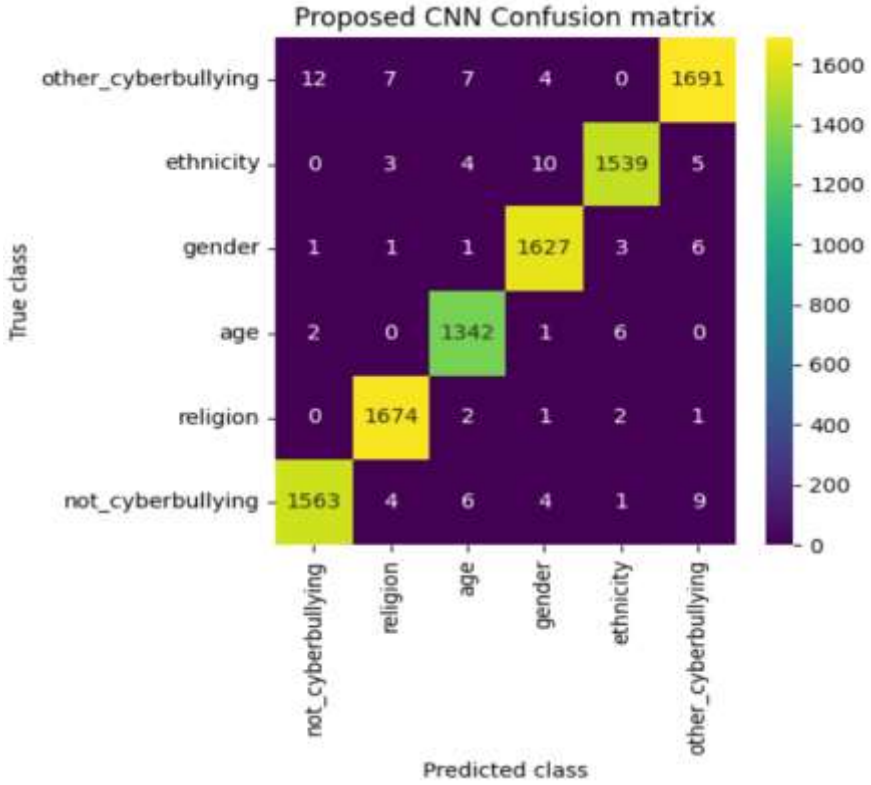

Fig.7: Proposed CNN Confusion Matrix.

Metric	Existing DNN	Proposed CNN
Accuracy	88.77%	98.92%
Precision (Macro)	88.95%	98.91%
Recall (Macro)	88.89%	98.93%
F1-Score (Macro)	88.77%	98.92%
Sensitivity	96.69%	99.74%
Specificity	97.88%	100.00%

Table 1: Comparison for Existing DNN and Proposed CNN

The comparison in Table 1 highlights the performance differences between the Existing DNN and Proposed CNN models across various metrics. The Existing DNN achieves an accuracy of 88.77%, while the Proposed CNN significantly outperforms it with an accuracy of 98.92%, a difference of 10.15%. In terms of precision (macro average), the DNN scores 88.95%, whereas the CNN scores 98.91%, showing a 9.96% improvement in correctly identifying positive instances across classes. The recall (macro average) for the DNN is 88.89%, compared to 98.93% for the CNN, indicating a 10.04% better ability of the CNN to capture all relevant instances. The F1-score (macro average) mirrors this trend, with the DNN at 88.77% and the CNN at 98.92%, a 10.15% improvement, reflecting a balanced improvement in precision and recall. Sensitivity, which measures the true positive rate for the positive class, is 96.69% for the DNN and 99.74% for the CNN, a 3.05% increase, showing the CNN's superior ability to detect cyberbullying instances. Specificity, which measures the true negative rate, is 97.88% for the DNN and a perfect 100.00% for the CNN, a 2.12% improvement, indicating the CNN's flawless performance in identifying non-cyberbullying instances.

5. CONCLUSION

The bot cyberbullying classification application developed demonstrates a robust system for detecting and categorizing various types of cyberbullying in social media text, specifically tweets. By leveraging a structured pipeline that includes data preprocessing, feature extraction, and machine learning models, the application effectively processes raw tweet data to identify instances of cyberbullying across multiple categories, such as religion, age, gender, ethnicity, and more.The implementation of two models—a Deep Neural Network (DNN) and a Convolutional Neural Network (CNN)—provides a comparative analysis of performance, with the CNN demonstrating superior results in terms of accuracy, precision, recall, and F1-score. A user-friendly GUI interface enhances interaction by enabling seamless dataset uploading, exploratory data analysis, model training, and prediction, making the tool accessible to users without deep technical expertise. Preprocessing steps—including text cleaning, tokenization, stopword removal, and stemming—ensure the data is well-prepared for model training. The use of TF-IDF vectorization and sequence padding caters to the specific requirements of the DNN and CNN models, respectively. The application's ability to save and load trained models adds efficiency for repeated usage, while detailed evaluation metrics provide transparency in model performance. Overall, this application serves as an effective tool for identifying cyberbullying, contributing to safer online environments through timely detection and intervention.

REFERENCES

[1] Abarna, S.; Sheeba, J.I.; Jayasrilakshmi, S.; Devaneyan, S.P. Identification of Cyber Harassment and Intention of Target Users on Social Media Platforms. Eng. Appl. Artif. Intell. 2022, 115, 105283. [PubMed]

[2] Giumetti, G.W.; Kowalski, R.M. Cyberbullying via SocialMedia and Well-being. Curr. Opin. Psychol. 2022, 45, 101314. [PubMed]

[3] Akram, W.; Kumar, R. A Study on Positive and Negative Effects of Social Media on Society. Int. J. Comput. Sci. Eng. 2017, 5, 351–354.

[4] Trotter, J.L.; Allen, N.E. The Good, the Bad, and the Ugly: Domestic Violence Survivors' Experiences with their Informal Social Networks. Am. J. Community Psychol. 2009, 43, 221–231. [PubMed]

[5] Rosa, H.; Pereira, N.S.; Ribeiro, R.; Ferreira, P.C.; Carvalho, J.P.; Oliveira, S.; Coheur, L.; Paulino, P.; Simão, A.V.; Trancoso, I. Automatic Cyberbullying Detection: A Systematic Review. Comput. Hum. Behav. 2019, 93, 333–345.

[6] Nandhini, B.; Sheeba, J. Online Social Network Bullying Detection Using Intelligence Techniques. Procedia Comput. Sci. 2015, 45, 485–492.

[7] Hemphill, S.A.; Kotevski, A.; Heerde, J.A. Longitudinal Associations Between Cyberbullying Perpetration and Victimization and Problem Behavior and Mental Health Problems in Young Australians. Int. J. Public Health 2015, 60, 227–237.

[8] Ioannou, A.; Blackburn, J.; Siringhini, G.; De Chrisiofaro, E.; Kouriellis, N.; Sirivianos, M.; Zaphiris, P. From Risk Factors to Detection and Intervention: A Metareview and Practical Proposal for Research on Cyberbullying. In Proceedings of the 2017 IEEE IST-Africa Week Conference, Windhoek, Namibia, 30 May–2 June 2017; pp. 1–8.

[9] Al Mazari, A. Cyber-Bullying Taxonomies: Definition, Forms, Consequences and Mitigation Strategies. In Proceedings of the 5th IEEE International Conference on Computer Science and Information Technology, Amman, Jordan, 27–28 March 2013; pp. 126–133.

[10] Perera, A.; Fernando, P. Accurate Cyberbullying Detection and Prevention on Social Media. Procedia Comput. Sci. 2021, 181, 605–611.

[11] Evangelio, C.; Rodriguez-Gonzalez, P.; Fernandez-Rio, J.; Gonzalez-Villora, S. Cyberbullying in Elementary and Middle School Students: A Systematic Review. Comput. Educ. 2022, 176, 104356.

[12] Schneider, S.K.; O'Donnell, L.; Stueve, A.; Coulter, R.W. Cyberbullying, School Bullying, and Psychological Distress: A Regional Census of High School Students. Am. J. Public Health 2012, 102, 171–177.

[13] Sharif, O.; Hoque, M.M. Tackling Cyber-Aggression: Identification and Fine-Grained Categorization of Aggressive Texts on Social Media Using Weighted Ensemble of Transformers. Neurocomputing 2022, 490, 462–481.

[14] Isaza, G.; Munõz, F.; Castillo, L.; Buitrago, F. Classifying Cybergrooming for Child Online Protection Using Hybrid Machine Learning Model. Neurocomputing 2022, 484, 250–259.

[15] Ramana, D.; Reddy, T.H. Detection of Online Hate in Social Media Platforms for Twitter Data: A Prefatory Step. In Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications, Aizawl, India, 25–26 June 2021; pp. 411–419.

[16] Roy, P.K.; Bhawal, S.; Subalalitha, C.N. Hate Speech and Offensive Language Detection in Dravidian Languages Using Deep Ensemble Framework. Comput. Speech Lang. 2022, 75, 101386.