

An Explainable, Context-Aware AI Framework for Cyber-Aggression Detection and Behavioural Profiling in British English Social Media Discourse

¹D.V. Sateesh Kaladhar Reddy, ²Dr. S. B. Sowjanya Kumar, ³S. Natesan, ⁴B. Murali Prasad

^{1,4}Assistant Professor, CSE Department, Sri Venkatesa Perumal College of Engineering & Technology, Andhra Pradesh, India

²Professor, CSE Department, Sri Venkatesa Perumal College of Engineering & Technology, Andhra Pradesh, India

³Associate Professor, CSE Department, Sri Venkatesa Perumal College of Engineering & Technology, Andhra Pradesh, India

Abstract - In today's digital era, the escalating phenomenon of cyberbullying is a pervasive and growing concern. With the increasing prevalence of social media platforms, such as Twitter, online abusive behaviour has become a significant issue that often leads to unpleasant experiences for users. Manual detection of abnormal and bullying behaviour within the realm of social media is inherently not scalable [1]. Moreover, most existing studies on cyberbullying detection have been predominantly conducted in English and very limited work has been done on (a widely used language in Asia). This paper presents an approach for detecting cyberbullying in Roman Urdu tweets and identifying abuser profiles on Twitter. Firstly, we develop a text corpus of Roman Urdu tweets with user profile data. Subsequently, we employ Gated Recurrent Unit (GRU) model coupled with the application of word2vec technique for word embedding to develop a cyberbullying detection model [3]. Furthermore, we present temporal abusive tweet probability analysis method to provide announced analysis of the number of bullying and non-bullying tweets sent by individuals within a specific time interval. To evaluate the performance, we compare the GRU-based approach with other machine learning models.

Keywords: Gated Recurrent Unit, Term Frequency Inverse Document Frequency (TF-IDF), Deep learning, Cyberbullying, XGBoost, Boosting Classifier (GBM) and Support Vector Machine (SVM).

I. INTRODUCTION

The most creative and challenging phase of the life cycle is system and design. The term design describes a final system and the process by which it is developed. It refers to the technical specifications that will be applied in implementation the candidate system [17]. The design may be defined as "the process of applying various techniques and principles for the

purpose of defining a device, a process or a system in sufficient details to permit its physical realization".

The design's goal is how the output is to be produced and in what format samples of the output and input are also presented. Second input data and database files have to be designed to meet the requirements of the proposed output. The processing phase is handled through the program construction and testing. Finally, details related to justification of the system and an estimate of the impact of the candidate system on the users and the organization are documented and evaluated by management as a step toward implementation.

The importance of software design can be stated in a single word "Quality". Design provides us with representation of software that can be assessed for quality. Design is the only way that we can accurately translate a customer's requirements into a finished software product or system without design we risk building an unstable system, that might fail if small changes are made or may be difficult to test, or one who's quality can't be tested [19]. So, it is an essential phase in the development of a software product.

II. LITERATURE SURVEY

2.1 Cyber bullying Detection and Hate Speech Identification using Machine Learning Techniques

Bullying has been prevalent since the beginning of time, it's just the ways of bullying that have changed over the years, from physical bullying to cyberbullying. According to Williard (2004), there are eight types of cyberbullying such as harassment, denigration, impersonation, etc. It's been around 2 decades since social media sites came into the picture, but there haven't been a lot of effective measures to curb social bullying and it has become one of the alarming issues in recent times. Our paper presents an analytical review of cyberbullying detection approaches and assesses methods to recognize hate speech on social media [5]. We aim to apply

traditional supervised classification methods as well as some novel ensemble machine learning techniques using a manually annotated open-source dataset for this purpose. This paper does a comparative study of various Supervised algorithms, including standard, as well as ensemble methods. The evaluations of the result based upon the scores obtained by accuracy shows that Ensemble supervised methods have the potential to perform better than traditional supervised methods.

2.2 Cyber bullying detection on social media using machine learning

Usage of internet and social media backgrounds tends in the use of sending, receiving and posting of negative, harmful, false or mean content about another individual which thus means Cyberbullying. Bullying over social media also works the same as threatening, calumny, and chastising the individual. Cyberbullying has led to a severe increase in mental health problems, especially among the young generation [9]. It has resulted in lower self-esteem, increased suicidal ideation. Unless some measure against cyberbullying is taken, self-esteem and mental health issues will affect an entire generation of young adults. Many of the traditional machine learning models has been implemented in the past for the automatic detection of cyberbullying on social media. But these models have not considered all the necessary features that can be used to identify or classify a statement or post as bullying [13]. In this paper, we proposed a model based on various features that should be considered while detecting cyberbullying and implement a few features with the help of a bidirectional deep learning model called BERT.

2.3 Aggression detection in social media from textual data using deep learning models

With the advancement of technology, social media such as Facebook, Twitter, etc. plays an important role in communication whether it is texting, sharing photos, audio-video calls or expressing views through comments. Along with these advantages, it has some negative sides as well which brings aggression towards some section of people. Such aggression, hatred in social media needs to be detected and prevented automatically which is the main objective of our work. We have worked on Hindi, English and Hindi-English (code-mixed) datasets. We used features like word vectors, aggressive words (manually created dictionary), sentiment scores, parts of speech and emojis for the classification task [17]. We experimented with several machine learning and deep learning models and the results indicate that XGBoost Classifier, Gradient Boosting Classifier (GBM) and Support Vector Machine (SVM) are most suited for the task.

Therefore, the output of the three classifiers was used for majority voting which provides f-scores of 68.13, 54.82 and 55.31 for the English, Hindi and code-mixed datasets respectively.

2.4 DEA-RNN: A hybrid deep learning approach for cyber bullying detection in Twitter social media platform

Cyberbullying (CB) has become increasingly prevalent in social media platforms. With the popularity and widespread use of social media by individuals of all ages, it is vital to make social media platforms safer from cyberbullying. This paper presents a hybrid deep learning model, called DEA-RNN, to detect CB on Twitter social media network. The proposed DEA-RNN model combines Elman type Recurrent Neural Networks (RNN) with an optimized Dolphin Echolocation Algorithm (DEA) for fine-tuning the Elman RNN's parameters and reducing training time [23]. We evaluated DEA-RNN thoroughly utilizing a dataset of 10000 tweets and compared its performance to those of state-of-the-art algorithms such as Bi-directional long short-term memory (Bi-LSTM), RNN, SVM, Multinomial Naive Bayes (MNB), Random Forests (RF). The experimental results show that DEA-RNN was found to be superior in all the scenarios. It outperformed the considered existing approaches in detecting CB on Twitter platform. DEA-RNN was more efficient in scenario 3, where it has achieved an average of 90.45% accuracy, 89.52% precision, 88.98% recall, 89.25% F1-score, and 90.94% specificity.

2.5 Offensive language detection on social media based on text classification

There is a concerning rise of offensive language on the content generated by the crowd over various social platforms. Such language might bully or hurt the feelings of an individual or a community. Recently, the research community has investigated and developed different supervised approaches and training datasets to detect or prevent offensive monologues or dialogues automatically. In this study, we propose a model for text classification consisting of modular cleaning phase and tokenizer, three embedding methods, and eight classifiers. Our experiments show a promising result for detection of offensive language on our dataset obtained from Twitter. Considering hyperparameter optimization, three methods of AdaBoost, SVM and MLP had highest average of F1-score on popular embedding method of TF-IDF.

2.6 Pioneer dataset and automatic recognition of Urdu handwritten characters using a deep autoencoder and convolutional neural network

Automatic recognition of Urdu handwritten digits and characters, is a challenging task. It has applications in postal address reading, bank's cheque processing, and digitization and preservation of handwritten manuscripts from old ages. While there exists a significant work for automatic recognition of handwritten English characters and other major languages of the world, the work done for Urdu language is extremely insufficient. This paper has two goals. Firstly, we introduce a pioneer dataset for handwritten digits and characters of Urdu, containing samples from more than 900 individuals. Secondly, we report results for automatic recognition of handwritten digits and characters as achieved by using deep auto-encoder network and convolutional neural network [18]. More specifically, we use a two-layer and a three-layer deep autoencoder network and convolutional neural network and evaluate the two frameworks in terms of recognition accuracy. The proposed framework of deep autoencoder can successfully recognize digits and characters with an accuracy of 97% for digits only, 81% for characters only and 82% for both digits and characters simultaneously.

III. EXISTING SYSTEM

- Social media becomes the virtual medium for bullying, shielding the bully's identity, making detecting cyber bullying a complex and challenging mission to protect online communities.
- Cyberbullying cases increase with volumized Internet usage because it can be easily committed anonymously, leading to a grave public health concern that brings many negative impacts, such as mental, psychological, and social problems [5].
- While cyberbullying victims end to suffer from mental health problems such as depression, anxiety, loneliness, and anhedonia, some are reported to be committing self-injurious behaviour and suicidal ideation.

3.1 Disadvantages of Existing System

- The existing system offers unsatisfactory accuracies.
- Trolling account doesn't have a real account to fool such systems or experts who can change writing styles and behaviours so that no patterns are found.

IV. PROPOSED SYSTEM

- In this system is to automate the detection of cyberbullying instances from the unruly post, deeming the problem as a text classification task with the help of state-of-the-art techniques using artificial intelligence and natural language processing knowledge.

- By natural language processing, text classification is frequently employed in identifying the category of a given corpus through several stages, such as text preprocessing, feature extraction, and the development of a classification model [24].

4.1 Advantages of Proposed System

- Detection cyberbullying attacks efficiently.
- Our system improves the accuracy of detection.

4.2 Modules

- Service Provider
- Train and Test
- Remote User
- Graphical Analysis

V. SYSTEM ARCHITECTURE

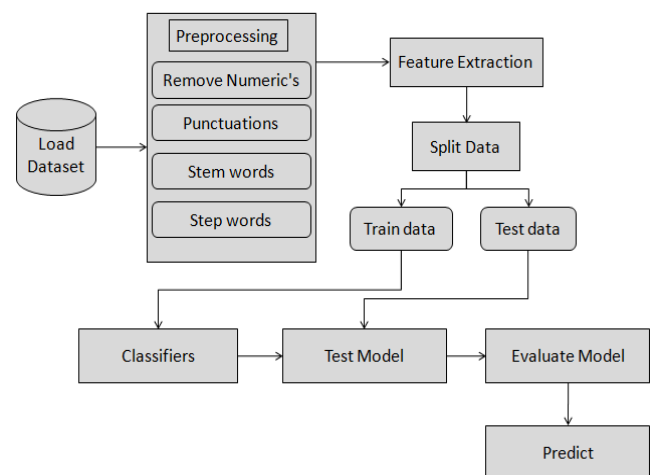


Figure 5.1: System Overview

VI. IMPLIMENTATION

6.1 Input design

Input design includes data mediums used for inputting data and validations that are to be done during data entry. Different messages regarding data are given to guide users during data entry. Validation checks are done for each input.

Data entry screens are designed so that the system interacts with the user in providing an effective dialogue. Fields in the screen are logically arranged to help the user.

The design is the process of converting the user-originated inputs into a computer-based format. The goal of the input design is to make the data entry easier, logical and

free from error. Errors in the input data are controlled by input design [14].

The application has been developed in a user-friendly manner.

The windows have been designed in such a way that during the processing the cursor is placed in the position where the data must be entered. If any of the data going into the system is wrong then the process and output will magnify these errors.

The decisions made during design of input are:

- 1) To achieve the highest possible level of accuracy.
- 2) To provide a list of possible choices and help while accepting the input for an important field wherever possible outputs from computer system are required primarily to communicate the results of processing to the users. They are also used to provide a permanent copy of these results for later consultation/verification [19].

6.2 Output Design

Output refers to the results and information that are generated by the system. Output is the main reason for developing the system and based on this, the usefulness and applicability of system are evaluated.

Outputs from computer systems are required primarily to communicate the results of processing to users. Efficiently designed outputs enhance the understandability of the information.

According to the requirements of the system, various types of outputs are considered and designed as follows.

Internal outputs, whose destination is within the organization and which require careful design because they are the user's main interface with the computer. Interactive outputs, in which the user communication with the computer is essential [20].

6.3 Module Description

Service Provider Module

- ✓ In this module, service provider login to the system using valid username and password. After login successful, he can perform the following operations like train and test dataset, view trained and tested accuracy and view remote users.

Train and Test Model

- ✓ In this module, the service provider split the Used dataset into train and test data of ratio 70 % and 30 % respectively. The 70% of the data is consider as train data which is used to train the model and 30% of the data is consider as test which is used to test the model

Remote User

- ✓ In this module, the remote user register to the system, and login to the system valid username, and password. After login successful, he can perform view profile, predict the abuser profile.

Graphical Analysis

- ✓ In this module, display the graphs like accuracy and predicted ratio of the system. Various factors take into consideration for the graph analysis. In this phase plot the charts like bar chart and so others.

6.4 Algorithms

Support Vector Machine

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. An SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

Logistic Regression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Decision tree

- Trees are constructed through an algorithmic approach that identifies ways to split the data set based on different conditions.

- It is one of the most widely used practical methods for supervised learning.
- These are non-parametric method used for both classification and regression

Gradient Boosting Classifier

- Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in

which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.

6.5 Test Case

TC. No	Test Case	Input	Expected Output	Observed Output	Result
1	Login	Enter Wrong User Name and Password	Invalid Login Details	User name and Password are invalid	Pass
2	Login	Enter User Name and Password	Login Successful	Login Successful	Pass
3	Mobile Number	Enter Alphanumeric characters	Mobile number must be digits only	Mobile number in 10 digits only	Fail
4	Upload file	Browse file	file uploaded successfully	file Uploaded successful	Pass

VII. EXPERIMENTAL RESULTS



Figure 7.1: Home Page

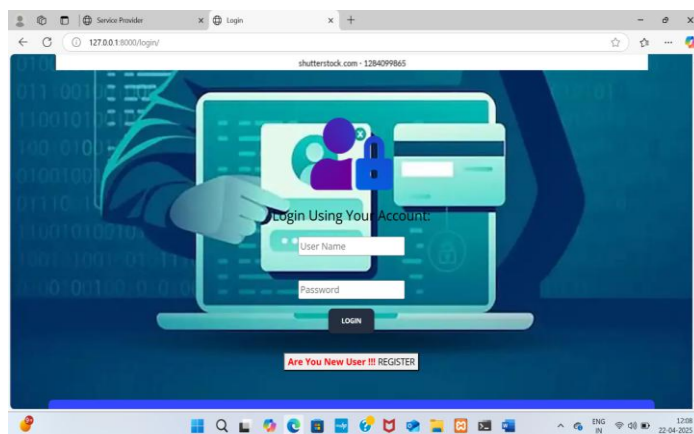


Figure 7.3: User Login



Figure 7.2: Admin Login

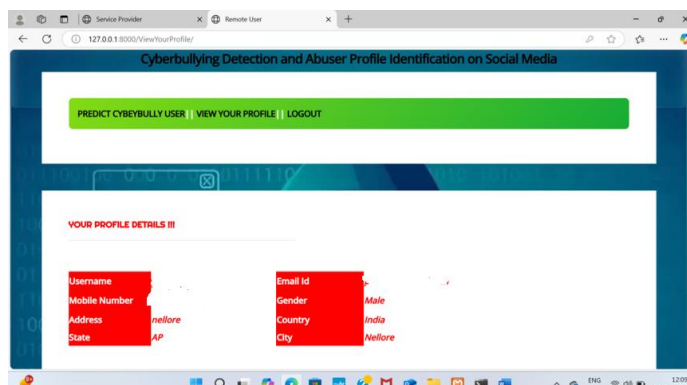


Figure 7.4: View Profile

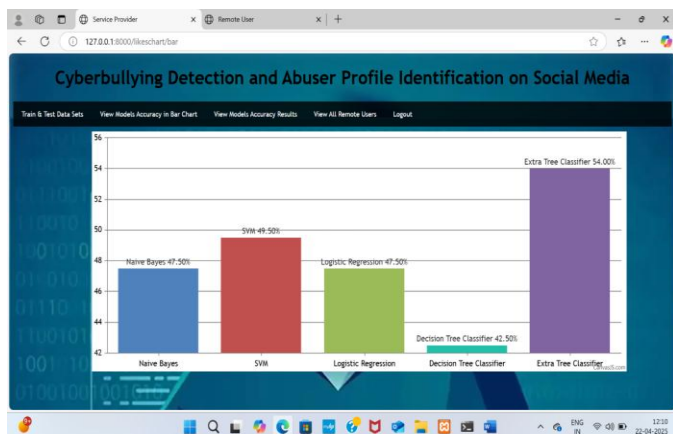


Figure 7.5: View Trained and Tested Accuracy Results

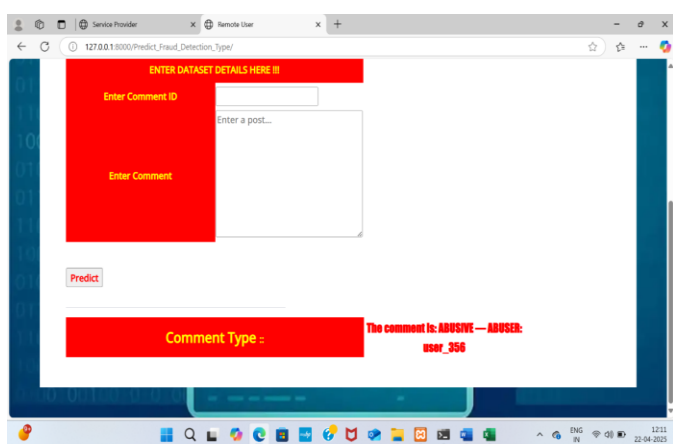


Figure 7.6: Predict the comment type

VIII. CONCLUSION

The findings of this study underscore the critical challenges faced by of cyberbullying content detection, primarily stemming from the limited availability of comprehensive datasets containing user profile information. This scarcity imparts significant implications for the accurate identification and categorization of abusive content spanning diverse cyberbullying categories in the Roman Urdu text on social media. The proposed method, utilizing the GRU model with pre-learned embeddings, exhibited remarkable advancements over conventional machine learning and deep learning models, including LSTM, Bi-LSTM with attention layers, and CNN [18]. These advancements were observed across key performance metrics such as accuracy, precision, and F-measure, with the GRU-based method attaining an accuracy rate of 97%, outperforming alternative methods in this domain. Furthermore, our investigation highlighted the effectiveness of employing Decision Tree and Gradient Boosting as conventional machine learning classifiers in conjunction with TF-IDF for cyberbullying detection.

IX. FUTURE ENHANCEMENT

Future enhancements of Cyberbullying Detection and Abuser Profile Identification on social media could involve integrating advanced deep learning models for more accurate context understanding. Real-time detection systems can be developed to intervene promptly and prevent harm. Cross-platform analysis could help identify repeat offenders across different social networks. Improved multilingual support would allow detection in diverse languages and dialects.

REFERENCES

- [1] M. Woodward. (Jan. 2026). Social Media in Pakistan—2023 Stats Platform Trends. Accessed: Apr. 15, 2023. [Online]. Available: <https://oosga.com/social-media/pak/>
- [2] T. Agrawal and V. D. Chakravarthy, “Cyberbullying detection and hate speech identification using machine learning techniques,” in *Proc. 2nd Int. Conf. Interdiscipl. Cyber Phys. Syst. (ICPS)*, May 2024, pp. 182–187.
- [3] D. Aditya, S. Kalaskar, O. Kumbhar, and R. Dhumal, “Cyber bullying detection on social media using machine learning,” in *Proc. ITM Web Conf.*, vol. 40, 2023, pp. 30–38, doi: 10.1051/itmconf/20214003038.
- [4] U. Khan, S. Khan, A. Rizwan, G. Atteia, M. M. Jamjoom, and N. A. Samee, “Aggression detection in social media from textual data using deep learning models,” *Appl. Sci.*, vol.12, no.10, p. 5083, May 2022.
- [5] B. A. H. Murshed, J. Abawayj, S. Mallappa, M. A. N. Saif, and H. D. E. Al-Araki, “DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform,” *IEEE Access*, vol. 10, pp. 25857–25871, 2023.
- [6] P. Hajibabaei, M. Malekzadeh, M. Ahmadi, M. Heidari, A. Esmaeilzadeh, R. Abdolazimi, and J. H. J. Jones, “Offensive language detection on social media based on text classification,” in *Proc. IEEE 12th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2022, pp. 92–98.
- [7] H. Ali, A. Ullah, T. Iqbal, and S. Khattak, “Pioneer dataset and automatic recognition of Urdu handwritten characters using a deep autoencoder and convolutional neural network,” *Social Netw. Appl. Sci.*, vol. 2, no. 2, pp. 1–12, Feb. 2021.
- [8] H. Ali, K. Iqbal, G. Mujtaba, A. Fayyaz, M. F. Bulbul, F. W. Karam, and A. Zahir, “Urdu text in natural scene images: A new dataset and preliminary text detection,” *Peer J. Comput. Sci.*, vol. 7, p. e717, Sep. 2021.
- [9] F. Rasheed, M. Anwar, and I. Khan, “Detecting cyberbullying in Roman Urdu language using natural

- language processing techniques,” *Pakistan J. Eng. Technol.*, vol. 5, no. 2, pp. 198–203, Sep. 2022.
- [10] A.M. Alduailaj and A. Belghith, “Detecting Arabic cyberbullying tweets using machine learning,” *Mach. Learn. Knowl. Extraction*, vol. 5, no. 1, pp. 29–42, Jan. 2023.
- [11] M. Bilal, A. Khan, S. Jan, and S. Musa, “Context-aware deep learning model for detection of Roman Urdu hate speech on social media platform,” *IEEE Access*, vol. 10, pp. 121133–121151, 2022.
- [12] B. A. Chandio, A. S. Imran, M. Bakhtyar, S. M. Daudpota, and J. Baber, “Attention-based RU-BiLSTM sentiment analysis model for Roman Urdu,” *Appl. Sci.*, vol. 12, no. 7, p. 3641, Apr. 2022.
- [13] R. Ghosh, B. T. Student, S. Nowal, and G. Manju, “Social media cyberbullying detection using machine learning in Bengali language,” *J. Int. J. Eng. Res. Technol.*, vol. 10, no. 5, pp. 190–193, 2021.
- [14] M. M. Khan, K. Shahzad, and M. K. Malik, “Hate speech detection in Roman Urdu,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 20, no. 1, pp. 1–19, Jan. 2021.
- [15] K. Sreelakshmi, B. Premjith, and K. P. Soman, “Detection of hate speech text in Hindi-English code-mixed data,” *Proc. Comput. Sci.*, vol. 171, pp. 737–744, Jan. 2020.
- [16] H. Rizwan, M. H. Shakeel, and A. Karim, “Hate-speech and offensive language detection in Roman Urdu,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Jan. 2020, pp. 2512–2522.
- [17] H. Nurrahmi and D. Nurjanah, “Indonesian Twitter cyberbullying detection using text classification and user credibility,” in *Proc. Int. Conf. Inf. Commun. Technol. (ICOIAC)*, Mar. 2018, pp. 543–548.
- [18] E. A. Abozinadah, A. V. Mbaziira, and J. H. J. Jones, “Detection of abusive accounts with Arabic tweets,” *Int. J. Knowl. Eng.*, vol. 1, no. 2, pp. 113–119, 2015.
- [19] Z. Zhang, D. Robinson, and J. Tepper, “Detecting hate speech on Twitter using a convolution-GRU based deep neural network,” in *The Semantic Web, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds., Cham, Switzerland: Springer*, 2018, pp. 745–760.
- [20] B. Bhatia, A. Verma, and R. Katarya, “Analysing cyberbullying using natural language processing by understanding jargon in social media,” in *Proc. ICSAC*, 2021, pp. 397–406.
- [21] J. J. Dooley, J. Pyzalski, and D. Cross, “Cyberbullying versus face-to-face bullying: A theoretical and conceptual review,” *J. Psychol. Psychotherapy*, vol. 9, no. 3, pp. 1–11, 2019.
- [22] J. W. Patchin and S. Hinduja, “Cyberbullying among adolescents: Implications for empirical research,” *J. Adolescent Health*, vol. 53, no. 4, pp. 431–432, Oct. 2013.
- [23] G. Sarna and M. P. S. Bhatia, “Content based approach to find the credibility of user in social networks: An application of cyberbullying,” *Int. J. Mach. Learn. Cybern.*, vol. 8, no. 2, pp. 677–689, Apr. 2017.
- [24] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, “Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth,” *Psychol. Bull.*, vol. 140, no. 4, p. 1073, 2014.
- [25] G. Fulantelli, D. Taibi, L. Scifo, V. Schwarze, and S. C. Eimler, “Cyber bullying and cyberhate as two interlinked instances of cyber-aggression in adolescence: A systematic review,” *Frontiers Psychol.*, vol. 13, May 2022, Art. no. 909299.

Citation of this Article:

D.V. Sateesh Kaladhar Reddy, Dr. S. B. Sowjanya Kumar, S. Natesan, & B. Murali Prasad. (2026). An Explainable, Context-Aware AI Framework for Cyber-Aggression Detection and Behavioural Profiling in British English Social Media Discourse. *International Current Journal of Engineering and Science (ICJES)*, 5(10), 1-7. Article DOI: <https://doi.org/10.47001/ICJES/2026.510001>
