IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 3, May-June 2025

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203118

Neural Network-Based Approach for Detecting Cyber Bullying on Social Platforms

N. Nikhil, V. Pranav, P. Shiva Sai Karthikeya, T.Devender Rao

Department of CSE, Guru Nanak Institutions Technical Campus, Hyderabad, India

ABSTRACT: The rise of Information and Communication Technologies has amplified social networking, but also increased the prevalence of cyber bullying. Manual, user-dependent mechanisms like reporting and blocking are often inefficient. This research proposes a Neural Network-based approach for detecting cyber bullying on social platforms, exploring both Conventional Machine Learning and Transfer Learning techniques. A comprehensive dataset with structured annotations was used. Features such as textual content, sentiment and emotional cues, static and contextual embeddings, psycholinguistics, term lists, and toxicity indicators were extracted. A key contribution is the introduction of toxicity features for cyber bullying detection. Among neural models, contextual Word Convolutional Neural Network (Word CNN) achieved a high F-measure. When combined in a Logistic Regression model, these features significantly improved performance, surpassing Linear SVC in training efficiency and high- dimensional feature handling. Transfer Learning using fine-tuned Word CNN further enhanced training speed. A Flask-based web application was developed to implement real-time detection, achieving strong accuracy.

I. INTRODUCTION

GENERAL

The rise of Information and Communication Technologies has amplified social networking, but also increased the prevalence of cyber bullying. Manual, user-dependent mechanisms like reporting and blocking are often inefficient. This research proposes a Neural Network-based approach for detecting cyber bullying on social platforms, exploring both Conventional Machine Learning and Transfer Learning techniques. A comprehensive dataset with structured annotations was used. Features such as textual content, sentiment and emotional cues, static and contextual embeddings, psycholinguistics, term lists, and toxicity indicators were extracted. A key contribution is the introduction of toxicity features for cyber bullying detection. Among neural models, contextual Word Convolutional Neural Network (Word CNN) achieved a high F-measure. When combined in a Logistic Regression model, these features significantly improved performance, surpassing Linear SVC in training efficiency and high- dimensional feature handling. Transfer Learning using fine-tuned Word CNN further enhanced training speed. A Flask-based web application was developed to implement real-time detection, achieving strong accuracy.

In recent years, hate speech has become increasingly widespread, affecting both face-to-face interactions and digital communication channels. This rise can be attributed to several interconnected factors. One significant contributor is the anonymity afforded by online platforms, particularly social media networks, which often encourages users to engage in more aggressive, unfiltered, and harmful expressions of opinion. Without fear of immediate social or legal consequences, individuals may feel emboldened to voice prejudiced views that they would otherwise suppress in offline settings. Additionally, the modern culture of open expression—where users frequently share personal beliefs, frustrations, and ideologies without restraint—has inadvertently contributed to the proliferation of hate speech. The viral nature of content on platforms such as Twitter, Facebook, and Instagram allows such messages to reach wide audiences rapidly, amplifying their social impact.

Deep Learning (DL), a powerful subfield of machine learning, enables the modeling of complex patterns in data, often through unsupervised learning using unlabeled datasets. In domains such as data mining and text classification, DL techniques have gained significant traction, offering promising results in tasks like hate speech detection and opinion classification. A variety of deep learning architectures have been applied in these contexts, including Feedforward Neural Networks, Deep Belief Networks, Convolutional Neural Networks (CNN), Restricted Boltzmann Machines, Recurrent Neural Networks (RNN), and Stacked Denoising Autoencoders.

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203118

SCOPE OF THE PROJECT

This study uses cutting-edge Natural Language Processing (NLP) and Deep Learning approaches to tackle the growing problem of cyberbullying in online plaforms. The scope includes integrating contextual embeddings using Word CNN and investigating innovative features, such as toxicity indications. The study explores the creation of effective classification algorithms that can recognize instances of cyberbullying in text-based online communication automatically.

OBJECTIVE

Examine and put into practice cutting-edge NLP and Deep Learning strategies to improve the identification of cyberbullying on online platforms. To increase the precision of cyberbullying detection algorithms, investigate the incorporation of novel variables, such as toxicity indicators, in addition to traditional textual and sentiment data. Create strong classification models that outperform traditional machine learning techniques by utilising Word CNN for contextual embeddings. In order to achieve high accuracy in real-time identification and prevention, integrate cyberbullying detection into a Flask online platform as a practical implementation of the created models.

II. PROBLEM STATEMENT

The increasing prevalence of cyber bullying on social media platforms necessitates the development of advanced and automated detection systems. Existing approaches rely heavily on user intervention through manual reporting or simplistic classifiers, which are often inefficient in identifying the nuanced and context- dependent nature of online harassment. This project addresses these limitations by proposing a neural network-based framework that leverages a wide range of features—textual content, sentiment cues, emotional indicators, static and contextual embeddings, psycholinguistic markers, and a newly introduced toxicity feature set. Central to this approach is the use of a Word Convolutional Neural Network (Word CNN), fine- tuned through transfer learning to adapt to the specific linguistic patterns associated with cyber bullying. The model demonstrates high performance in classification tasks, achieving a superior F-measure compared to traditional classifiers such as Linear SVC, while also offering faster training efficiency in handling high- dimensional data.

To support real-world deployment and accessibility, a Flask-based web application has been developed, enabling real-time cyber bullying detection with strong accuracy and responsiveness. The incorporation of transfer learning not only boosts performance but also reduces the computational burden of training from scratch. By introducing toxicity features and refining deep contextual embeddings, this system enhances the granularity with which harmful content can be detected, including subtle, indirect, or masked bullying. This comprehensive methodology not only surpasses current state-of-the-art models but also contributes to more ethical and safer digital interactions. The project stands out by combining performance optimization, interpretability, and scalability, making it a promising solution for modern-day social media moderation and digital well-being initiatives.

EXISTING SYSTEM

- The existing system focuses on addressing the resource-intensive nature of machine learning (ML) classifier training, particularly with the rising challenge posed by large datasets and the prevalence of Deep Neural Networks (DNN). Feature Density (FD) is analyzed as a means to estimate ML classifier performance before training, aiming to optimize the training process.
- The study underscores the environmental impact of resource-intensive training, especially concerning the escalating CO2 emissions associated with large-scale ML models. The research aims to minimize the demands for powerful computational resources and enhance efficiency in Natural Language Processing, with a specific emphasis on dialog classification, such as cyber bullying detection.

EXISTING SYSTEM DISADVANTAGES

- Mostly uses a density-based metric, which may exclude out subtle language elements that are important for detecting cyber bullying.
- Impairing the classifier's capacity to identify tiny signs of cyber bullying and intricate language patterns.
- It not being able to integrate sophisticated features and embeddings.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203118

III. LITERATURE SURVEY

Various approaches have been proposed for the detection of cyberbullying, focusing on both the analysis of content and the development of robust classification models. In the study by [8], a novel model is introduced that adopts a dual-definition framework for cyberbullying detection. The model combines a creative application of Convolutional Neural Networks (CNNs) for content analysis with a strategic mechanism aimed at mitigating classification inaccuracies. Compared to other existing methods, the model demonstrates improved accuracy and superior classification performance, highlighting the effectiveness of CNNs in processing and understanding social media content.

Another significant contribution is the systematic review conducted by [9], which analyzed 186 records retrieved from online research databases. From this pool, ten key literature reviews were selected to evaluate the role and effectiveness of machine learning (ML) in combating cyberbullying. The analysis reveals that most cyberbullying detection models primarily rely on content-based features extracted from social media posts. Commonly employed algorithms in this domain include Support Vector Machines (SVM), Naive Bayes, and Convolutional Neural Networks. These models generally utilize textual features to identify offensive or harmful language. The findings suggest that machine learning offers a promising avenue for cyberbullying prevention. ML not only supports the development of automated screening systems but also complements traditional adolescent education initiatives.

PROPOSED SYSTEM

- The automatic detection method of the proposed system tackles the issue of cyber bullying in social networks. The system uses a combination of textual, sentiment, emotional, static, and contextual information, using a big dataset and structured annotations. This method is distinct in that it incorporates toxicity factors to improve the identification of cyber bullying.
- When it comes to managing high-dimensionality features and training time, the system performs better than Linear SVC. When compared to base models, Transfer Learning enhances performance and speeds up training calculations by fine-tuning WORD CNN. Furthermore, a 97.06% accuracy rate in actual usage is guaranteed by a Flask web implementation.

PROPOSED SYSTEM ADVANTAGES

- Detects complex semantic connections in text.
- Enabling it to recognize patterns at various abstraction levels, which is useful for recognizing a variety of cyber bullying expressions.
- Improving cyber bullying detection without compromising performance by using pre- trained models and speeding up training.

APPLICATION GENERAL

The proposed neural network-based cyber bullying detection system can be effectively integrated into social media platforms, messaging applications, and online forums to monitor user-generated content in real time. By automatically identifying abusive or harmful language, the system can assist in flagging and filtering cyber bullying content before it escalates, thereby protecting users—especially vulnerable groups like teenagers and young adults.

FUTURE ENHANCEMENT

This work is limited to binary text classification even though the cyber-bullying corpus encompasses input from other roles within cyberbullying episodes, but it does not help us to determine who posted the cyberbullying post. this work can be extended to classify the participant roles such as harassers, victims, bystanders, and non-bullies. Since the conversation was handled as an individual post, the research could be extended to account for the relationship between posts to capture the interaction between users within cyberbullying episodes.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203118

IV. CONCLUSION

In conclusion, the unanticipated rise in cyberbullying as a result of technology advancement has highlighted the pressing need for efficient preventive measures. Automated detection methods must be developed and put into place since they have the potential to have severe and broad effects on Internet users. This is a preventative measure that also makes a substantial contribution to reducing the number of cyberbullying incidences. Although textual characteristics have been the mainstay of past techniques for classifying cyberbullying, this research has taken a more thorough approach by exploring many feature categories. We have broadened the range of possible indications for cyberbullying detection by examining textual features, sentiment and emotional features, embeddings, psycholinguistic features, word lists characteristics, and toxicity features. Our models' use of Word CNN has shown to be quite successful, as seen by their remarkable 97.06% accuracy rate. This illustrates how reliable and effective the suggested method is in locating and stopping instances of cyberbullying. The model's excellent accuracy rate demonstrates its adaptability and recognition of many patterns and settings in the intricate world of online communication.

REFERENCES

[1]. B. Cagirkan and G. Bilek, ‘‘Cyberbullying among Turkish high school students,’’ Scandin. J. Psychol., vol. 62, no. 4, pp. 608–616, Aug. 2021, doi: 10.1111/sjop.12720.
[2]. P. T. L. Chi, V. T. H. Lan, N. H. Ngan, and N. T. Linh, ‘‘Online time, experience of cyber bullying and practices to cope with it among high school students in Hanoi,’’ Health Psychol. Open, vol. 7, no. 1, Jan. 2020, Art. no. 205510292093574, doi: 10.1177/2055102920935747.
[3]. A. López-Martínez, J. A. García-Díaz, R. Valencia-García, and A. Ruiz-Martínez, ‘‘CyberDect. A novel approach for cyberbullying detection on Twitter,’’ in Proc. Int. Conf. Technol. Innov., Guayaquil, Ecuador: Springer, 2019, pp. 109–121, doi: 10.1007/978-3-030- 34989-9_9.
[4]. R. M. Kowalski and S. P. Limber, ‘‘Psychological, physical, and academic correlates of cyberbullying and traditional bullying,’’ J. Adolescent Health, vol. 53, no. 1, pp. S13–S20, Jul. 2013, doi: 10.1016/j.jadohealth.2012.09.018.
[5]. Y.-C. Huang, ‘‘Comparison and contrast of piaget and Vygotsky’s theo-ries,’’ in Proc. Adv. Social Sci., Educ. Humanities Res., 2021, pp. 28–32, doi: 10.2991/assehr.k.210519.007.
[6]. A. Anwar, D. M. H. Kee, and A. Ahmed, ‘‘Workplace cyberbullying and interpersonal deviance: Understanding the mediating effect of silence and emotional exhaustion,’’ Cyberpsychol., Behav., Social Netw., vol. 23, no. 5, pp. 290–296, May 2020, doi: 10.1089/cyber.2019.0407.
[7]. D. M. H. Kee, M. A. L. Al-Anesi, and S. A. L. Al-Anesi, ‘‘Cyberbul-lying on social media under the influence of COVID-19,’’ Global Bus. Organizational Excellence, vol. 41, no. 6, pp. 11– 22, Sep. 2022, doi: 10.1002/joe.22175.
[8]. I. Kwan, K. Dickson, M. Richardson, W. MacDowall, H. Burchett, C. Stansfield, G. Brunton,
K. Sutcliffe, and J. Thomas, ‘‘Cyberbullying and children and young people’s mental health: A systematic map of systematic reviews,’’ Cyberpsychol., Behav., Social Netw., vol. 23, no. 2, pp. 72–82, Feb. 2020, doi: 10.1089/cyber.2019.0370.
[9]. R. Garett, L. R. Lord, and S. D. Young, ‘‘Associations between social media and cyberbullying: A review of the literature,’’ mHealth, vol. 2, p. 46, Dec. 2016, doi: 10.21037/mhealth.2016.12.01.
[10]. M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, and K. Araki, ‘‘Automatic extraction of harmful sentence patterns with application in cyberbullying detection,’’ in Proc. Lang. Technol. Conf. Poznań, Poland: Springer, 2015, pp. 349–362, doi: 10.1007/978-3-319-93782-3_25.
[11]. M. Ptaszynski, P. Lempa, F. Masui, Y. Kimura, R. Rzepka, K. Araki, M. Wroczynski, and
G. Leliwa, ‘‘"Brute-force sentence pattern extortion from harmful messages for cyberbullying detection,’’’
J. Assoc. Inf. Syst., vol. 20, no. 8, pp. 1075–1127, 2019.
[12]. M. O. Raza, M. Memon, S. Bhatti, and R. Bux, ‘‘Detecting cyber-bullying in social commentary using supervised machine learning,’’ in Proc. Future Inf. Commun. Conf. Cham, Switzerland: Springer, 2020, pp. 621–630.
[13]. D. Nguyen, M. Liakata, S. Dedeo, J. Eisenstein, D. Mimno, R. Tromble, and J. Winters, ‘‘How we do things with words: Analyzing text as social and cultural data,’’ Frontiers Artif. Intell., vol. 3, p. 62, Aug. 2020, doi: 10.3389/frai.2020.00062.
[14]. J. Cai, J. Li, W. Li, and J. Wang, ‘‘Deeplearning model used in text classification,’’ in Proc. 15th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP), Dec. 2018,
pp. 123–126, doi: 10.1109/ICCWAMTIP.2018.8632592.
[15]. N. Tiku and C. Newton. Twitter CEO: We Suck at Dealing With Abuse. Verge. Accessed: Aug. 17, 2022. [Online]. Available: https://www.theverge.com/2015/2/4/7982099/twitter-ceo- sent-memo-taking- personal-responsibility-for-the.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com