Design and Implementation of an AI-Based Cyberbullying Detection System

Nitya Shree R, Divyashree S, Neha G, Pooja Kulkarni, Poornima K
Department of Information Science and Engineering
RNS Institute of Technology, Bengaluru

ABSTRACT: Growing up in a world of instantaneous communication, contemporary "digital natives" have developed limitless relationships. Cyberbullying, or "wilful and repeated injuries caused by the use of electronic devices," has become more prevalent because of this quick evolution, which frequently makes it difficult to distinguish between harmful and acceptable behaviours This study suggests a method for automatically identifying bullying traces on social networks by utilising natural language processing and machine learning. Using semantic and syntactic sentence attributes, our method effectively clusters documents with bullying content using Growing Hierarchical SOMs. The model was tested on YouTube and Formspring, but it was refined for Twitter. The outcomes show that this unsupervised approach performs well in detecting cyberbullying in a range of contexts.

I.INTRODUCTION

A crucial component of text mining is text categorisation, which includes the classification of documents using supervised, unsupervised, or semi-supervised methods. Machine learning techniques such as Decision Trees, K-Nearest Neighbours, Naïve Bayes, Support Vector Machines, and N-grams have mechanised this process, which was formerly done by hand. Due to budget constraints, few research examine Bangla text, despite the fact that it is commonly used in English. Supervised learning models, such as Decision Trees, KNN, Naïve Bayes, and SVM, are useful for detecting cyberbullying. Three stages are involved: pre-processing (such as stemming, tokenisation, and stop-word removal), feature extraction (statistical methods for pertinent attributes), and classification (using algorithms to group text into categories). These techniques can categorise high-dimensional, sparse, and noisy text using social media data sites, making it possible to identify instances of cyberbullying effectively and promoting safer online spaces.

II.LITERATURE SURVEY

It examines many analyses and tests that have been oversee in the field of interest, a literature review is crucial. It examines the results that have already been made public while accounting for the scope of the project and other project features. The primary objective of a literature review is to conduct a comprehensive examination of the project's past, identifying shortcomings in the current configuration and emphasizing issues that still require attention. In addition to illuminating the project's past, the topics covered also draw attention to the problems and weaknesses that drove the project's inception and suggested solutions.

Jun,2023

Cyberbullying is a distressing type of online wrongdoing that frequently manifests as hurtful language on social media. To solve this, intelligent systems are required for automatic detection. Although conventional models for machine learning have been utilised to solve this problem, current research has highlighted deep learning methods, that have demonstrated improved results. This study assesses the efficacy of four deep learning models—BLSTM, GRU, LSTM, and RNN—for detecting insults in social commentary. Data preprocessing steps included text cleaning, tokenization, stemming, lemmatization, and stop-word removal.

The findings indicated that the BLSTM model beat RNN, LSTM, and GRU regarding precision and F1-measure scores. This research not only highlights the power of deep learning models within cyberbullying detection, but also suggests the promise of future hybrid solutions to tackle this severe menace.

Dec,2021

Social media has developed into an indispensable tool for communication, collaboration, and idea exchange. However, the anonymity it provides has resulted in an increase in hate speech and cyberbullying, a growing concern that has piqued the interest of scholars worldwide. While much of the study focuses on mature languages, there is a notable gap in studies that address resource-poor languages, particularly Roman Urdu, which is extensively spoken in South Asia. To close this gap, we performed significant preprocessing on Roman Urdu microtext, including constructing a slang-phrase lexicon and removing domain-specific stop words to minimize corpus dimensionality. We used advanced models such as RNN-LSTM, RNN- BLSTM, and CNN, with different epochs, layers, and hyperparameters. Our findings revealed that the RNN-LSTM and RNN-BLSTM models outperformed the others in detecting cyberbullying, with validation accuracy of 85.5% and 85%, respectively, and F1 scores of 0.7 and 0.67.

III. PROPOSED SYSTEM

Natural language processing (NLP) and machine learning (ML) are used in the suggested automated cyberbullying detection method to find offensive content in online chats. In order to improve accuracy, it collects information from public databases, messaging apps, and social media. It also handles noise, emojis, and multilingual, unstructured content. The density of "bad" words and the "badness" of sentences—which is based on how serious the offending keywords are—are important characteristics for detection. Additional attributes are also considered, including the frequency of exclamation or question marks (which indicate emotional intensity) and the density of capital letters (which indicates yelling). Because Support Vector Machines (SVM) are resilient when working with data that is not balanced, they are utilised. Clean input is ensured by preprocessing techniques like tokenisation, lowercase conversion, and stop word removal, which allow messages to be classified as bullying or non-bullying, with detection that is scalable in different platforms.

IV. ADVANTAGES OF PROPOSED SYSTEM

1. Real-Time Scalability: The system processes large data volumes in real-time, enabling quick cyberbullying detection across platforms, making it both scalable and cost-effective.

2. Contextual Detection: Using advanced NLP and machine learning techniques like sentiment analysis, sarcasm detection, BERT, and LSTM, the system accurately identifies subtle cyberbullying forms, including sarcasm and irony.

3. Continuous Learning: Regular feedback and retraining improve the system's accuracy, adapting to new language trends and bullying tactics, ensuring consistent and objective detection.

V. METHODOLOGY

Discuss the machine learning or algorithms for deep learning that were chosen for the study, including Transformer-based models (e.g., BERT), Support Vector Machines (SVM), Random Forests, Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). Training and Testing: Describe how the models are trained and tested, including the cross-validation approach, the train-test split, and any methods (such SMOTE or class weighting) that are employed to resolve class imbalance. Custom Features: Explain how any custom features (such as the frequency of derogatory terms or the usage of abusive language) were developed and how they relate to identifying cyberbullying. Validation: Describe the metrics used to assess performance (e.g., accuracy, precision, recall, F1-score, AUC-ROC) and the techniques used for model validation, such as cross-validation.

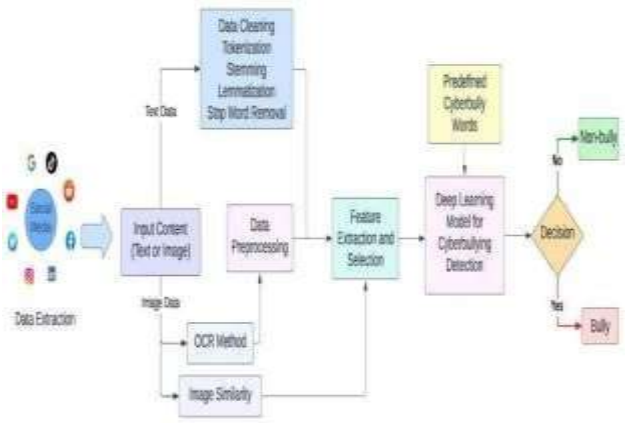
VI. SYSTEM DESIGN

Data Collection: Twitter's API with OAuth is used to securely retrieve tweets in real-time. In order to overcome issues like duplicate filtering and linguistic variety, the system collects tweet content, user information, and metadata from 20 accounts.

Storage: Enhanced by metadata such as timestamps and geolocation, tweets are kept in a relational database with unique identifiers (TwitterId, TwitterDesc, UserId) for effective searching and analysis.

Preprocessing Stopwords: To cut down on noise and enhance analysis, stopwords are eliminated. For domain-specific and multilingual data, custom stopword lists are utilised to optimise processes such as sentiment analysis.

Data Cleaning: Stopwords are eliminated, text is normalised (lowercased, punctuation removed), and missing data is handled. For efficient model training, the cleaned dataset is organised with distinct IDs and tweet content.



VII. RESULTS

Results from the suggested LSTM-based model for cyberbullying identification are encouraging, indicating that it works well on a variety of social media sites. The model's accuracy rates on Facebook, Instagram, and Twitter datasets were 91.26%, 94.49%, and 96.64%, respectively. These findings demonstrate the model's high reliability in identifying cases of cyberbullying and its ability to analyse textual material effectively. The benefits of utilising deep learning methods such as LSTM, which can more successfully capture context and sequential linkages in textual data, are highlighted by the higher accuracy rates as compared to previous approaches. This study lays a strong foundation for the creation of automated methods to counteract online harassment moreover validating the viability of employing such sophisticated algorithms for cyberbullying identification.

VIII. SNAPSHOTS

Fig 8.1 Register page

The registration screen, where users can establish an account by entering information like their username and password, is shown in the picture.

Fig8.2 Login page

The screenshot shows the login page with the password and username fields that guarantee safe user authentication.

Fig8.3 User home page

Reverse chronologically arranged tweets with text, timestamps, and engagement indicators like likes and comments are displayed on the user home page. An input box allows users to create new tweets, promoting a simplified, content-focused experience for smooth communication.

Fig 8.4 Blocked account page

According to the screenshot of the prohibited account page, the reason the account has been blocked is has more than five bullying remarks. It notifies the user that the account was reported for breaking harassment-related platform rules.

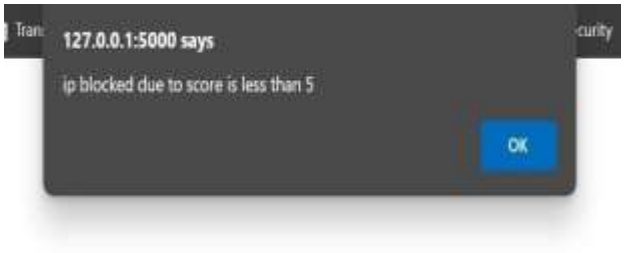


Fig 8.5 Notification raised

The screenshot highlights the infraction by displaying a warning notification that the user's account has been blocked because they've shared more than five bullying comments.

IX. CONCLUSION

In conclusion, this study tackles the problem of cyberbullying by utilising sophisticated feature extraction methods and techniques for machine learning, such Support Vector Machine (SVM) and k-Nearest Neighbours (KNN). The work captures subtle textual patterns by carefully preparing the data and experimenting with models such as Bag of Words, TF-IDF, and Word2Vec. The detection accuracy is increased by combining KNN's flexibility to local patterns with SVM's capacity to manage intricate relationships. This method contributes create a safer online environment by offering insightful information about cyberbullying detection. In addition to increasing technological correctness, the research fosters a socially conscious online environment, lessens the effects of cyberbullying, and increases empathy and inclusivity in online groups.

Ethical considerations, such as privacy, fairness, and bias, must also be prioritized to ensure responsible system deployment. Ultimately, AI and machine learning possess the capacity to greatly combat cyberbullying, providing safer online environments and empowering individuals to manage their digital interactions effectively.

X. FUTURE ENHANCEMENT

The future of AI and machine learning in cyberbullying detection holds promising advancements. AI will integrate multimodal sentiment analysis, analyzing text, voice, and images to detect bullying in various formats. Advanced Natural Language Processing (NLP) will improve context understanding, allowing AI to differentiate between harmful and harmless interactions. Real-time detection will enable proactive intervention, flagging bullying content as it's created. Personalized models will adapt based on factors like age, culture, and relationships, improving detection accuracy across settings like schools and workplaces. Emotion recognition and psychological profiling will assess the emotional impact of bullying, analyzing facial expressions, tone, and physiological responses. AI systems will continuously learn, reducing erroneous negative and positive results, and recognizing emerging bullying tactics. Ethical considerations such as privacy protection and fairness will ensure AI systems are culturally sensitive and unbiased. These advancements will help AI both detect and prevent cyberbullying, creating safer online space.

XI. REFERENCE

[1] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, ‘‘Cyberbully ing detection solutions based on deep learning architectures,’’ Multimedia Syst., vol. 29, no. 3, pp. 1839–1852, Jun. 2023, doi: 10.1007/s00530-020 00701-5.

[2] A. Dewani, M. A. Memon, and S. Bhatti, ‘‘Cyberbullying detection: Advanced preprocessing techniques & deep learning architecture for Roman Urdu data,’’ J. Big Data, vol. 8, no. 1, pp. 1–20, Dec. 2021, doi: 10.1186/s40537-021- 00550-7.

[3] M. Di Capua, E. Di Nardo and A. Petrosino, Unsupervised cyberbullying detection in social networks, ICPR, pp. 432-437,doi:10.1109/ICPR.2016.7899672. (2016)

[4] Mandal, Ashis Kumar,Rikta Sen. "Supervised learning methods for Bangla web document categorization." International Journal of Artificial Intelligence & Applications, IJAIA, Vol 5, pp. 5, 10.5121/ijaia.2014.5508

[5] Dani Harsh, Jundong Li, and Huan Liu, “Sentiment Informed Cyberbullying Detection in Social Media” Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Spinger, Cham, 2017.

[6] Dinakar, Karthik, Roi Reichart, and Henry Lieberman. “Modeling the detection of Textual Cyberbullying.” The Social Mobile Web 11.02(2011):11-17

[7] A. saravanaraj, J. I. sheebaassistant, S. Pradeep, and D. Dean, “Automatic Detection of Cyberbullying From Twitter.” IRACST-International J. Comput. Sci. Inf. Technol. Secur., vol. 6, no. 6,pp. 2249-9555,2016.

[8] S. Hnduja and J. W. Patchin “Cyberbullying: Identification, Prevention, & Response,”Cyberbullying Res. Cent, no. October, pp. 1-9,2018

[9] S. Salawu, Y. He, and J. LUMSDEN, “Approaches to Automated Detection of Cyberbullying: A Survey,” IEEE Trans. Affect. Compute., pp. 1–25, 2017

[10] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, ―Mean birds: Detecting aggression and bullying on Twitter,‖ WebSci 2017 - Proc. 2017 ACM Web Sci. Conf., pp. 13–22, Jun. 2017, doi: 10.1145/3091478.3091487.

[11] M. Anul Haq, ―DBoTPM: A Deep Neural Network-Based Botnet Prediction Model,‖ 2023, doi: 10.3390/electronics12051159

[12] M. M. Kermani and R. Azarderakhsh, ―Reliable Architecture-Oblivious Error Detection Schemes for Secure Cryptographic GCM Structures,‖ IEEE Trans. Reliab., vol. 68, no. 4, pp. 1347–1355, 2019, doi: 10.1109/TR.2018.2882484