Robertanet: Enhancedroberta Transformer Based Model for Cyber Bullying Detection with Glove Features

Aneena Anto[1], Ann Mariya Joju[2], Ansel Shanavas[3], Anu Krishna K[4], Akhila V A[5]

[1,2,3,4]Student, Department of Computer Science and Engineering, IES College of Engineering, Thrissur, Kerala, India

[5] Assistant Professor, Department of Computer Science and Engineering, IES College of Engineering, Thrissur, Kerala, India

Email_id: aneenaanto05@gmail.com, annmariyajoju26@gmail.com, anselks450@gmail.com, anuk13899@gmail.com, akhilava@iesce.info

Abstract

Social media has become a vital platform for communication, but it also encourages harmful behaviours such as cyberbullying, trolling, and hate speech. Manual moderation is slow and expensive, making automatic detection essential. This work presents an enhanced RoBERTa transformer model combined with GloVe word embeddings to detect cyberbullying in tweets. The system is tested against various machine learning and deep learning models and achieves about 95% accuracy, along with high precision, recall, and F1 scores. Cross-validation further confirms its reliability. The results show that advanced transformer models supported by effective feature representation can provide a strong solution for detecting cyberbullying on social media. Beyond academic evaluation, such a system can be valuable in protecting vulnerable users, especially teenagers, from psychological harm. Moreover, the approach has potential to be adapted across multiple social platforms, making it a promising step towards safer digital communities.

Keywords— Human Trafficking, Missing Children, Facial Recognition, Deep Learning, VGG-Face, KNN Classifier, CCTV Surveillance, NLP, Real-Time Detection, AI in Law Enforcement

DOI: https://doi.org/10.5281/zenodo.18538516

1. Introduction

Cyberbullying has emerged as one of the most destructive consequences of the social media revolution, posing serious threats to the safety, dignity, and mental health of millions of users across the globe. Social networks such as Facebook, Twitter, Instagram, and LinkedIn have transformed the way people communicate, share knowledge, and build communities, but their openness and anonymity also make them vulnerable to misuse.

Among the most alarming out comes is cyberbullying, which involves the use of abusive language, insults, threats, or targeted harassment intended to embarrass, intimidate, or harm individuals. Unlike traditional bullying, which is often confined to physical settings, cyberbullying is pervasive, continuous, and amplified by the global reach of online platforms.

Manual moderation is impractical given the speed and scale of online interactions, where millions of posts and comments are generated every minute, making automated detection the only viable solution.

1.1 Overview

This project focuses on developing an advanced system for the automatic detection of cyberbullying on social media platforms. With the increasing use of online networks, harmful behaviours such as trolling, harassment, and hate speech have become widespread, creating serious risks for vulnerable users. Manual moderation of such content is slow, expensive, and impractical given the volume of data generated every second.

1.2 Significance of Study

The increasing use of social media has amplified the risk of harmful behaviours such as cyberbullying, trolling, and hate speech, making it essential to develop effective detection mechanisms. This study is significant because it addresses one of the most urgent challenges of the digital age: protecting users, especially young and vulnerable groups, from the negative psychological and social impacts of online harassment.

1.3 User Interface (UI)

A user-friendly web and mobile interface that allows donors to browse charity projects, make donations, and track the impact of their contributions. Charities will use the same interface to create fundraising campaigns and provide updates on how funds are being used.

1.4 Audit Module

An integrated audit system that periodically checks the consistency and accuracy of transactions recorded on the blockchain. This module ensures that all data is authentic and aligns with the reported activities of the charities.

2. Literature Survey

The issue of social media cyber bullying, notably Twitter (now referred to as X) and Facebook, is of significant concern due to its profound impact on users' well-being, especially among younger demographics who frequently use these platforms [19]. Ottosson [20] introduced a large language model (LLM) aimed at detecting cyberbullying on social media platforms. Utilizing the GPT-3 LLM, the study sought to minimize the gap in platform moderation. The outcomes indicate that the proposed model performs comparably to the preceding models.

The researchers developed a CNN-attention framework that amalgamated an attention layer with a convolutional pooling layer, enabling efficient extraction of cyberbullying-related keywords from users' tweets. The study utilized two sets of combinations. Initially, they combined CNN and ML models where convolutional layers served as feature extractors, and ML models like RF and LR were used for classification. In the subsequent structure, they employed combinations like CNN-XGB and CNN-LSTM for classification. The findings revealed that the proposed CNN-Attention framework out performed other learning models, achieving an impressive accuracy of 97.10%.

Wang et al. [22] presented a graphical convolutional method for underlying cyberbullying detection. The frame work leverages a semi-supervised online dynamic query expansion (DQE) process to automatically generate balanced data. This process extracts additional natural data points of a specific class from Twitter. They also introduced a graph convolutional network (GCN) classifier that operates on a graph, constructed from thresholded cosine similarities between tweet embeddings. The performance of this system was compared against different machine learning models coupled with various embedding techniques. The results show that for the proposed SOSNet, using SBERT, an accuracy score of 92.70% was achieved. In a separate study, Qudah et al. [23] suggested an improved system for cyberbullying detection utilizing an adaptive external dictionary (AED). The authors employed ML models such as RF, XGB, and CatBoost, and introduced ensemble voting models. The findings suggest that the proposed ensemble voting model, when used with AED, provide superior accuracy in detecting cyberbullying incidents.

Muneeretal. [28] proposed a modified BERT and stacking ensemble model to identify social media cyberbullying. A continuous bag of words(CBOW),alongwithaword2vec like feature extraction method is employed to establish weights in the embedding layer. An outstanding accuracy of 97.4% was achieved using the stacking ensemble learning method as revealed by experiment results for discovering cyberbullying on the tweet dataset. A related study used CBOW feature extraction and attention mechanism based on deep learning focussed on detecting cyberbullying. Various deep learning models, including Conv1DLSTM, LSTM, CNN, BiLSTM_Pooling, BiLSTM, and gated recurrent unit (GRU), were employed. Results underlined the dominance of the attention-based Conv1DLSTM classifier over the other applied approaches, achieving the highest accuracy of 94.49%. The existing approaches are marked by several limitations. First, feature engineering approaches are not very well studied in the context of machine learning models for cyberbullying detection. Secondly, the imbalanced class distribution is not investigated with respect to its impact on classification accuracy. Finally, BERT and its variants are not widely studied for the topic at hand. This study aims to fill this gap by proposing the RoBERTa model and investigating several feature engineering approaches.

3. Review of Methodology

3.1 System Design:

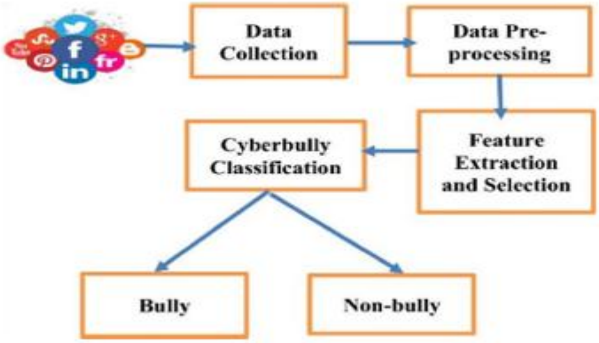

Figure 1: System Design

The cyberbullying classification pipeline begins with comprehensive data collection from various social media platforms such as Facebook, Twitter, and LinkedIn. These platforms serve as rich sources of user-generated content, where instances of cyberbullying often occur in the form of comments, posts, or direct messages. The collected data is typically unstructured and noisy, necessitating a robust pre-processing stage. This includes tasks such as tokenization, stop-word removal, normalization, and handling of emojis or slang, all of which help standardize the input for downstream Language analysis.

Once the data is cleaned, the next phase involves feature extraction and selection. This step is crucial for identifying linguistic and behavioural patterns that distinguish bullying content from benign interactions. Techniques such as TF-IDF, word embeddings, and syntactic parsing are employed to convert textual data into numerical representations. Feature selection algorithms then refine these representations by isolating the most informative attributes, which may include sentiment polarity, frequency of abusive terms, or contextual cues. In some implementations, metadata such as user history, posting time, and engagement metrics are also incorporated to enrich the feature space. This multi-dimensional approach improves the classifier's ability to detect nuanced forms of aggression.

The final stage of the pipeline is cyberbully classification, where machine learning or deep learning models are trained to categorize the data into ”Bully” or ”Non-bully” classes. Algorithms such as Support Vector Machines, Random Forests, or transformer-based models like RoBERTaNET are commonly used for this task. The classification outcome is then visualized or stored for further action, enabling moderators or automated systems to flag harmful content. Post-classification, feedback loops can be introduced to retrain the model with newly labeled data, enhancing adaptability over time. Additionally, explainability modules such as attention heatmaps or SHAP values can be integrated to interpret model decisions, fostering transparency and trust in automated moderation systems.

4. Review Of Datasets

A review of datasets for a blockchain-based transparent charity application ensures that the data used supports the system's goals of transparency and trust while being comprehensive and accurate.

4.1 Donation Data

The Donation Data comprises both transaction records and donation history. Transaction records capture essential details such as donor identities, donation amounts, dates, and the recipient charities. These records are important in maintaining a detailed and immutable account of all transactions, thus enhancing transparency and facilitating thorough audits. Ensuring the accuracy and completeness of these records, along with their seamless integration into blockchain technology is crucial for preserving their immutability. Additionally, donation history tracks past donations, providing valuable insights for trend analysis and evaluating the long-term impact of contributions. Maintaining consistency and historical accuracy in this dataset is vital for its effectiveness.

4.2 Charity Information

The dataset includes both charity profiles and project details. Charity profiles offer critical insights such as mission statements, operational details, and funding needs, which lend context and legitimacy to the organizations and assist donors in making informed decisions. The accuracy and completeness of these profiles are crucial, as they must be regularly updated and aligned with the blockchain system to maintain their relevance. Additionally, project details provide specific information about individual charity projects and their associated funding goals, ensuring donors are well-informed about how their contributions are utilized. The emphasis here is on ensuring that these details are both accurate and clearly communicated, with a strong integration into the donation tracking system to maintain transparency and accountability.

4.3 Donor Information

These datas consists of donor profiles and donor history. Donor profiles contain personal information such as names, contact details, and preferences, which help personalize the donor experience and facilitate communication. Ensuring data protection and privacy is paramount. Donor history includes records of past donations and interactions, which help understand donor behaviour and preferences. Accuracy and integration with current donation records are critical for a complete view of donor activity.

4.4 Smart Contract Specifications

Involves contract terms and contract performance data. Contract terms include the conditions and rules encoded in smart contracts, which automate and enforce donation processes and fund allocation. The accuracy and functionality of these contracts are crucial. Contract performance data consists of logs of smart contract executions, tracking how contracts are executed and ensuring compliance with their terms. This dataset must accurately reflect performance and highlight any issues or anomalies.

4.5 Financial Data

Financial data includes fund allocation and expense tracking. Fund allocation data details how funds are distributed among projects or purposes, ensuring that donations are used as intended and supporting financial transparency. Accuracy in these records is crucial for reporting and accountability. Expense tracking data provides information on charity expenditures, including administrative and project costs, helping to understand how funds are spent and ensuring consistency with allocated amounts.

4.6 System Performance Data

Performance data encompasses transaction times and error logs. Transaction times data measures how quickly transactions are processed, which is important for assessing system efficiency. Error logs document system errors and issues, helping identify and resolve problems to maintain smooth operation.

4.7 User Interaction Data

The dataset includes information on dashboard usage and feedback/support requests. Dashboard usage data captures how users engage with their dashboards, offering insights into usability, user behaviour, and identifying areas that may require enhancement. Meanwhile, feedback and support requests provide crucial details about user concerns and experiences, enabling the system to be refined and improved based on real-world user interactions. Both aspects of this dataset are vital for optimizing the user experience and ensuring the system meets user needs effectively.

5. Implementation Of Cyberbullying Detection and Classification

Implementation of the proposed cyberbullying detection and classification system is guided by well-defined hardware, software, functional, and non-functional requirements to ensure accurate detection, high performance, and reliability across multiple social media platforms.

5.1. Hardware Requirements:

The hardware configuration ensures efficient processing of large datasets and smooth execution of deep learning models for cyberbullying detection:

- Processor: Intel Pentium Core i3 or higher for effective text and image data processing.
- Primary Memory: Minimum 4GB RAM to handle model training and real-time data analysis.
- Storage: At least 320GB hard disk for storing datasets, model checkpoints, and logs.
- Additional Requirements: Stable internet connectivity for dataset access, cloud integration, and model deployment.

5.2. Software Requirements:

The software stack supports the end-to-end development, training, and deployment of the detection model:

- Operating System: Windows 8 or higher for compatibility with development and AI tools.
- Front-End Development: HTML and CSS for creating an intuitive and responsive user interface.
- Back-End Development: Python (Flask/Django) for implementing detection logic and MySQL for storing user and comment data.
- Machine Learning Frameworks: TensorFlow and PyTorch for training deep learning models such as CNNs, RNNs, or transformers (e.g., BERT, RoBERTa).
- Development Environment: Jupyter Notebook, Anaconda, or Visual Studio Code for efficient coding, testing, and debugging.

5.3. Functional Requirements:

The system incorporates essential functionalities to enable efficient detection, classification, and reporting of cyberbullying incidents:

User Registration and Authentication:

- Secure registration and login for users, moderators, and administrators.
- Role-based authentication to maintain privacy and restrict unauthorized access.

Data Collection and Preprocessing:

- Automated extraction of text, images, and comments from social media platforms.
- Preprocessing steps such as tokenization, stop-word removal, and stemming for text normalization.

Cyberbullying Detection and Classification:

- Machine learning and deep learning models trained to identify offensive, bullying content.
- Classification of detected content into categories such as harassment, hate speech, or insult.

Real-Time Monitoring and Alerts:

- Continuous monitoring of user interactions and instant detection of cyberbullying content.
- Automatic alert generation for moderators and reporting features for users.

Security and Privacy:

- Data encryption and secure API communication to prevent data breaches.
- Ethical data handling ensuring anonymity and compliance with privacy regulations.

Administrative Functions:

- Dashboard for administrators to monitor flagged content and manage user reports.
- Regular report generation on detection accuracy, user activity, and system performance.

Scalability and Accessibility:

- Support for large-scale social media data and multilingual inputs.
- Compatibility across platforms and devices for global accessibility.

5.4 Non-Functional Requirements:

The non-functional requirements of the RoBERTaNET system focus on ensuring its quality, efficiency, and adaptability. The system should be scalable to handle large and continuously growing social media datasets while maintaining high performance. It must provide robustness against noisy, imbalanced, and dynamic data by consistently producing reliable results. Efficiency in both training and inference is essential, with optimized use of computational resources for faster execution. The system should also be flexible enough to adapt to different platforms, languages, and datasets, ensuring wider applicability. Additionally, it must maintain reliability and deliver consistent accuracy across multiple runs.

6. Result And Discussion

Cyberbullying detection continues to be a pressing concern in online communication, as the increasing prevalence of social media exposes users to harmful interactions. The detection task is inherently challenging because online language often includes context-dependent expressions, sarcasm, slang, and long-range dependencies, which traditional machine learning models struggle to interpret accurately. While conventional deep learning techniques have im proved performance compared to basic methods, they often fail to capture subtle semantic and contextual cues in text. Transformer-based architectures, such as BERT and RoBERTa, have emerged as powerful tools for understanding language at a deeper level.

Their ability to model relationships between words across long sequences enables the detection of nuanced patterns that simpler models might overlook. These models represent a significant advancement in natural language processing, providing a more effective foundation for identifying cyberbullying content.

Ultimately, this comprehensive groundwork paves the way for developing a reliable, efficient, and practical cyberbullying detection system. The insights gained from this study will also contribute to ongoing research in online safety, offering guidance for future improvements and applications in monitoring and moderating online content.

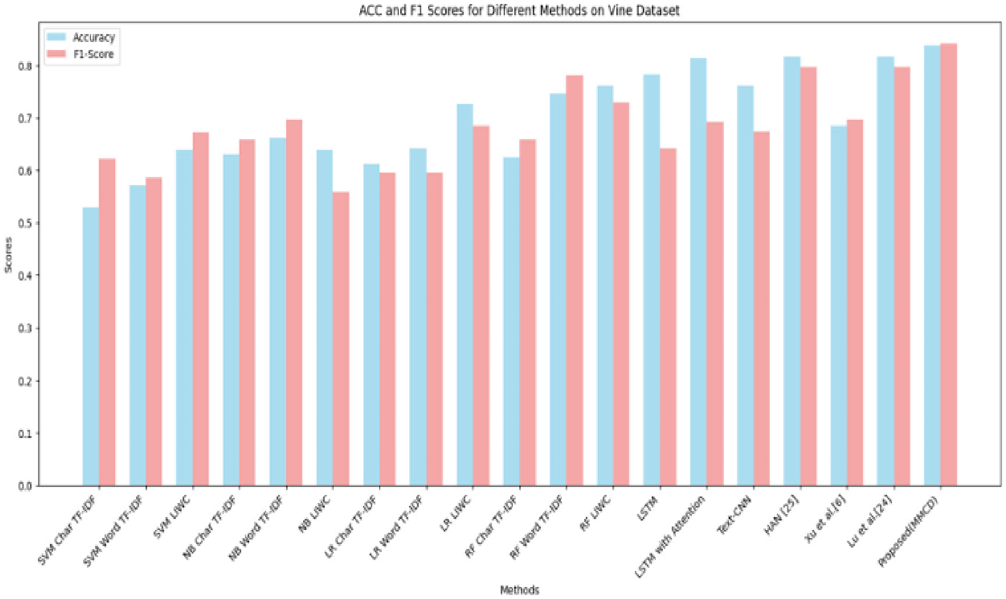

Figure 2: F1 Score

7. References

[1]. S. R. Sahoo and B. B. Gupta, “Classification of various attacks and their defence mecha nism in online social networks: A survey,” Enterprise Information Systems, vol. 13, no. 6, pp. 832–864, Jul. 2019.

[2]. S. Neelakandan, R. Annamalai, S. J. Rayen, and J. Arunajsmine, “Social media networks owing to disruptions for effective learning,” Procedia Computer Science, vol. 172, pp. 145–151, Sep. 2020.

[3]. M. Fire, G. Katz, and Y. Elovici, “Strangers intrusion detection detecting spammers and fake profiles in social networks based on topology anomalies,” Human Journal, vol. 1, no. 1, pp. 26–39, 2012.

[4]. B.Dean, “HowManyPeopleUseSocialMediain2021,”Backlinko. [Online]. Available: https://backlinko.com/social-media-users

[5]. S. Agrawal and A. Awekar, “Deep learning for detecting cyberbullying across multiple social media platforms,” in Proc. European Conference on Information Retrieval, 2018, pp. 141–153.

[6]. N. Selwyn, “Social media in higher education,” The Europa World of Learning, vol. 1, no. 3, pp. 1–10, 2012.

[7]. J.F.HairandM.Sarstedt,“Data,measurement,andcausalinferencesinmachinelearning: Opportunities and challenges for marketing,” Journal of Marketing Theory and Practice, vol. 29, no. 1, pp. 65–77, Jan. 2021.

[8]. D.Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, “Mean birds: Detecting aggression and bullying on Twitter,” in Proc. ACM Web Science Conf., Jun. 2017, pp. 13–22.

[9]. D.Ottosson,CyberbullyingDetectiononSocialPlatformsUsingLargeLanguageModels, 2023. [Online]. Available: https://www.diva-portal.org

[10]. A.AlhloulandA.Alam,BullyingTweetsDetectionUsingCNN-attention,SSRN4338998, 2023. Department of

Computer Science and Engineering 30 Cyberbullying Detection

[11]. J. Wang, K. Fu, and C.-T. Lu, “SOSNet: A graph convolutional network approach to f ine-grained cyberbullying detection,” in Proc. IEEE Int. Conf. Big Data, Dec. 2020, pp. 1699–1708.

[12]. H.Qudah, M.A.Alhija, and H.Tarawneh, “Improving Cyberbullying Detection Through Adaptive External Dictionary in Machine Learning,” 2023. [Online].

[13]. S. A. Mathur, S. Isarka, B. Dharmasivam, and J. C. D., “Analysis of tweets for cyberbul lying detection,” in Proc. 3rd Int. Conf. Secure Cyber Computing and Communications (ICSCCC), May 2023, pp. 269–274.

[14]. B. George Bokolo and Q. Liu, “Cyberbullying detection on social media using machine learning,” in Proc. IEEE Conf. Computer Communications Workshops (INFOCOM WKSHPS), May 2023, pp. 1–6.

[15]. M. Nisha and J. Jebathangam, “Detection and classification of cyberbullying in social media using text mining,” in Proc. 6th Int. Conf. Electronics, Communication and Aerospace Technology, 2022, pp. 856–861.