NLP Techniques Cyberbullying Text Analysis on Twitter

Rishi R
B Tech CSE
Department of Computing Technologies
SRM Institute of Science and Technology,
Kattankulathur Campus
Chengalpattu, Tamil Nadu
India
rr5299@srmist.edu.in

Mohammed Irfan M A
B Tech CSE
Department of Computing Technologies
SRM Institute of Science and Technology,
Kattankulathur Campus
Chengalpattu, Tamil Nadu
India
mm5610@srmist.edu.in

Dr.G.Balamurugan
Assistant Professor
Department of Computing Technologies
Faculty of Engineering and Technology
SRM Institute of Science and Technology,
Kattankulathur Campus
Chengalpattu, Tamil Nadu
India
balamurg1@srmist.edu.in

2024 10th International Conference on Communication and Signal Processing (ICCSP) | 979-8-3503-5306-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCSP60870.2024.10544177

Abstract **With the pervasive nature of social media, the rise of cyberbullying has become a critical concern, necessitating advanced methodologies for timely identification and analysis. This research delves into the application of Natural Language Processing (NLP) mechanisms to mention the challenges of cyberbullying within the dynamic and fast-paced environment of Twitter. The study explores the utilization of state-of-the-art NLP algorithms and methodologies for the classification and analysis of cyberbullying text. Techniques such as sentiment analysis, text categorization, and semantic understanding are employed to discern the nuanced and context-dependent nature of cyberbullying content on the Twitter platform. We present a comprehensive examination of various machine learning models, including but not limited to recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer methods like BERT, for their effectiveness in accurately classifying cyberbullying text. Additionally, attention is given to feature engineering and pre-processing strategies tailored to the unique characteristics of Twitter data.**

Keywords: Natural Language Processing (NLP); Recurrent neural networks (RNN); Bidirectional Encoder Representations from Transformers (BERT)

I. INTRODUCTION

Natural Language Processing (NLP) has emerged as an important field in the world of artificial intelligence, contributing to various applications, including social media analysis. In this context, the study focuses on employing advanced NLP techniques for cyberbullying text analysis on Twitter, a platform notorious for its prevalence of online harassment. With the proliferation of social media, understanding and mitigating cyberbullying has become imperative.[1]. This research endeavors to harness the power of machine learning algorithms and linguistic analysis to dissect and interpret the nuances of language used in tweets, aiming to unveil patterns indicative of cyberbullying incidents. The main procedure of the approach lies in leveraging state-of-the-art deep learning models, such as recurrent neural networks and transformers models like Bidirectional Encoder Representations from Transformers (BERT), etc.[2] These models excel in processing sequential data and capturing intricate relationships within text, enabling a more nuanced understanding of context and sentiment. Additionally, natural language understanding techniques, including word embeddings and sentiment analysis, are applied to unravel the semantic layers of Twitter posts. [3] The integration of deep learning methodologies allows for the identification of subtle linguistic cues that may escape traditional rule-based systems, enhancing the precision and recall of cyberbullying detection. Furthermore, the study delves into the realm of feature engineering, exploring the efficacy of tokenization, part-of-speech tagging, and named entity recognition in enhancing the robustness of the analysis. By dissecting the syntactic and semantic structures of tweets, the model gains a richer contextual understanding, essential for distinguishing between genuine conversations and instances of cyberbullying. Evaluation metrics such as F1 score, precision, and recall are employed to quantify the performance of the developed system, providing insights into its efficacy in real-world scenarios. [4]. In conclusion, this research not only contributes to the advancement of NLP techniques but also addresses a pressing societal issue, fostering a safer online environment through data-driven cyberbullying detection on Twitter.

II. PROBLEM STATEMENT

Cyberbullying has become a pressing societal issue, with online platforms serving as breeding grounds for malicious behavior. Twitter, a popular microblogging platform, has witnessed a surge in cyberbullying incidents. This project aims to address the growing concern by leveraging Natural Language Processing (NLP) techniques for the analysis of cyberbullying text on Twitter. The goal is to develop an effective and efficient system that can automatically identify and categorize instances of cyberbullying, thereby contributing to the creation of a safer online environment. To address these issues, this research focuses on developing a comprehensive automated system utilizing Convolutional Neural Networks (CNNs) with Keras and TensorFlow.

The goal is to create a robust, efficient, and accurate solution that seamlessly identifies plant seedlings and detects diseases, enabling timely intervention for disease control and optimal growth.

III. RELATED WORK

Gada et al. [5] utilizes conventional Machine Learning and Natural Language Processing text classification models that do not take into account the semantics of sentences. The objective of this study is to address this limitation. To achieve this goal, we employ word2vec to train personalized word embeddings, upon which we construct our LSTM-CNN architecture. Subsequently, we train the model on this architecture. The proposed approach enhances the system's performance and efficiency. Iwendi et al. [6] proposes Bidirectional LSTM (BLSTM) model maintains two unique input and forward input states using two separate LSTM components. One LSTM processes the input sequence conventionally, while the other processes the input sequence in reverse order. This bidirectional approach aims to collect textual information from both previous and further inputs, potentially enhancing the model's understanding. While this method generally achieves faster learning compared to a unidirectional approach, its effectiveness can vary depending on the task at hand, as depicted in Figure 3 illustrating the architecture of the BLSTM model utilized. This introduced methodology results in an enhanced accuracy rate. Ghosh et al. [7] This study proposes deep learning-based solutions to address the increasingly prevalent issue of cyberbullying in the contemporary landscape of social media and digital connectivity. Early detection and identification of such occurrences hold the potential to mitigate the impact of this unethical behavior. To this end, we employ CNN, LSTM, and bidirectional LSTM (Bi LSTM) networks, capable of detecting and categorizing texts into six distinct cyberbully classes. Our training and testing procedure utilizes a dataset comprising 159k input examples, encompassing diverse texts representing both non-bullying and bullying sentiments. The findings reveal that the proposed deep learning models attain an overall test accuracy of 0.97, 0.931, and 0.96 using CNN, LSTM, and Bi LSTM networks, respectively, positioning Bi LSTM as a suitable choice for cyberbully detection purposes. This approach fortifies the network's resilience and enhances its overall effectiveness. Al-Hashedi et al. [8] proposes, two deep learning algorithms, specifically, LSTM, and BLSTM, were tested. Pre-processing steps, such as oversampling, were conducted on some chosen social media datasets. Four distinct word embedding models, comprising word2vec, GloVe, and ELMO, were investigated for feature representations. ELMO, which considers word context by extracting information from surrounding words, addresses certain limitations of pre-trained word embedding models. This approach enhances the accuracy of detection time. Daas et al. [9] proposes, Long Short-Term Memory (LSTM) model, a deep learning technique, is used to detect and mitigate cyberbullying incidents. Utilizing LSTM for feature extraction, model training, and analysis yields significantly improved results. The final assessment of this approach demonstrates that LSTM achieves an accuracy of 74.13%, surpassing previous models by a considerable margin. This method enhances effectiveness by reducing retrieval time.

IV. PROPOSED METHOD

The proposed approach involves using tweets as input for the text preprocess the Twitter data to remove noise, tokenize the text, and handle any special characters or emojis. Then, utilize BERT for contextual embedding of the text, capturing nuanced meanings and context-specific information crucial for understanding cyberbullying instances. The BERT embeddings can be fed into an LSTM-based classification model, which leverages the temporal dependencies within the text data to further refine the classification. This hybrid approach allows for the extraction of both semantic and sequential information, enabling more accurate detection of cyberbullying instances on Twitter.

In the implementation, fine-tuning the pre-trained BERT model on a labeled dataset specific to cyberbullying tweets is essential to adapt the model to the task at hand. Additionally, integrating techniques such as attention mechanisms within the LSTM architecture can help prioritize important words or phrases in the classification process. Furthermore, post-processing steps such as thresholding or ensemble methods can be employed to enhance the model's performance and mitigate false positives. Overall, by combining the strengths of BERT's contextual embeddings and LSTM's sequential modeling capabilities, this proposed method offers a robust framework for effectively analyzing cyberbullying text on Twitter, contributing to the development of more accurate and reliable detection systems. Figure 1 illustrates the BERT method’s workflow.

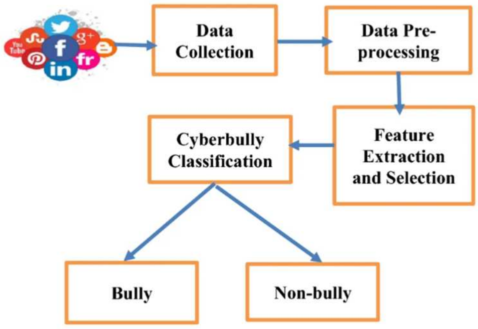


Fig. 1 Proposed Method

To analyze cyberbullying text on Twitter using LSTM (Long Short-Term Memory) neural networks, the proposed method would involve several key steps. Firstly, data collection would be imperative, gathering a sizable dataset of Twitter posts containing instances of cyberbullying, alongside non-cyberbullying tweets for comparison. Preprocessing techniques, such as tokenization, removing stopwords, and stemming or lemmatization, will be applied to clean and standardize the text data.

Following this, the dataset would be split into training, testing and validation sets, to evaluate and train the LSTM model effectively. Secondly, the LSTM model architecture is needed to be designed and then implemented. The LSTM model would comprise an embedding layer to convert words into dense vector representations, followed by one or more LSTM layers to capture the sequential nature of the text data. Additional layers, such as dropout layers, could be incorporated to prevent overfitting. During training, the model would learn how the tweets are classified as cyberbullying or non-cyberbullying based on the given input text data. Evaluation matrix such as accuracy, recall, precision, and F1-score would be used to assess this model's performance. Finally, the trained model would be implemented. Figure 2 illustrates the LSTM Architecture.

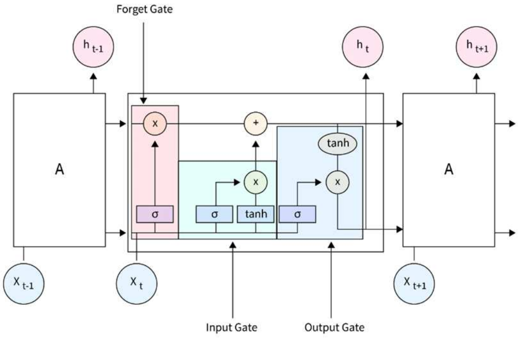


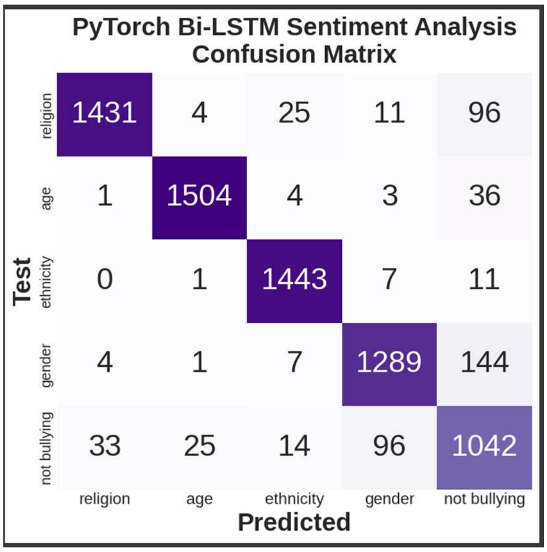
Fig. 2 LSTM Architecture

V. RESULT AND ANALYSIS

In this section, we examine the evaluation of the proposed approach's performance through confusion matrix. To assess the classification of BERT for this given dataset, we calculate the precision, F1 score, recall, support of the method, a dataset containing tweets is employed.

With 10 epochs of the cyberbullying detection model have an accuracy of about 95.50% which makes it a good model.

Fig. 3 shows the confusion matrix for the LSTM model where it has predicted measure in the X-axis and test method in the Y-axis. A Confusion matrix is an N x N matrix that is used for calculating the performance of a classification model, where N is the total number of target classes. The matrix compares the original target values with those values predicted by the deep learning prediction model. The performance scores of the algorithm are an accuracy of 92% and the F1 scores are over 91% for more densely populated classes .

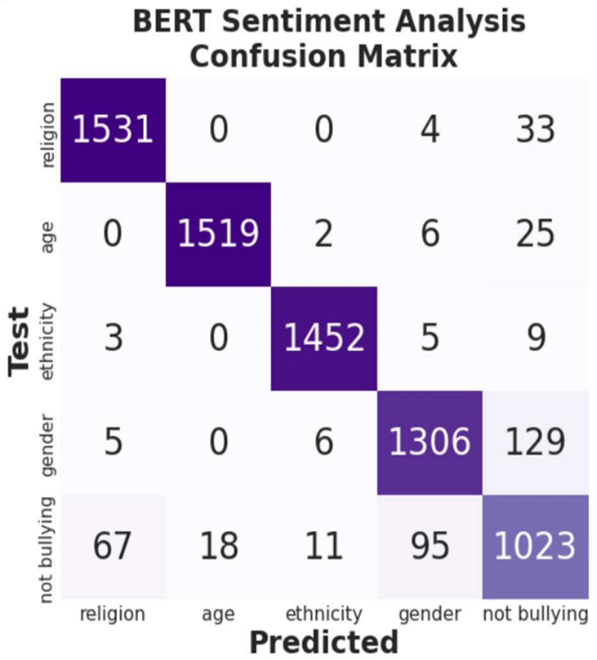


PyTorch Bi-LSTM Sentiment Analysis Confusion Matrix

Test \ Predicted	religion	age	ethnicity	gender	not bullying
religion	1431	4	25	11	96
age	1	1504	4	3	36
ethnicity	0	1	1443	7	11
gender	4	1	7	1289	144
not bullying	33	25	14	96	1042

Fig. 3 Confusion Matrix LSTM Model

Fig. 4 shows the confusion matrix for the BERT model, where it has predicted model in X-axis and test model in Y-axis . The performance matrix of BERT Classification model are quite higher, with an overall accuracy percent around 94.32% and F1 scores well over 94.91%.



BERT Sentiment Analysis Confusion Matrix

Test \ Predicted	religion	age	ethnicity	gender	not bullying
religion	1531	0	0	4	33
age	0	1519	2	6	25
ethnicity	3	0	1452	5	9
gender	5	0	6	1306	129
not bullying	67	18	11	95	1023

Fig. 4 Confusion Matrix BERT Model

Therefore, relative performance scores of BERT Classification model are quite higher than those resulted using the LSTM model, with an overall accuracy percent around 94.32% and F1 scores well over 94.91%.

VI. CONCLUSION

In the realm of cyberbullying text analysis on Twitter, the utilization of Natural Language Processing (NLP) techniques has become paramount in discerning and combating instances of online harassment. This conclusion delves into the efficacy of two prominent NLP models, Long Short-Term Memory (LSTM) and Bidirectional-Encoder Representations from Transformers (BERT), in their application to cyberbullying detection. Through a thorough examination of their performance metrics, it becomes evident that while both models exhibit proficiency in discerning abusive language, BERT outshines LSTM in terms of accuracy. LSTM, renowned for its ability to capture sequential dependencies, has long been a staple in NLP tasks, including sentiment analysis and language modeling. In the context of cyberbullying text analysis, LSTM effectively processes tweet sequences and identifies linguistic patterns indicative of online harassment. Its architecture, along with memory and gating mechanism, enables the model to retain contextual information over extended periods, thereby facilitating the detection of subtle nuances in abusive language. However, despite its commendable performance, LSTM may falter when faced with the complexity of Twitter data, which often includes slang, sarcasm, and evolving linguistic trends. These challenges may impede LSTM's accuracy and hinder its ability to effectively discern cyberbullying instances. On the contrary, BERT, a transformer-based model renowned for its bidirectional attention mechanism, emerges as a formidable contender in cyberbullying text analysis. By considering the entire context of a tweet and capturing bidirectional dependencies between words, BERT exhibits a remarkable ability to discern subtle nuances in language and identify instances of online harassment with unparalleled accuracy. Pretrained on vast corpora of text data, BERT encodes rich contextual information into its representations, enabling it to generalize effectively to diverse linguistic styles and nuances present in Twitter data. Moreover, BERT's adaptability and versatility make it well-suited for cyberbullying detection tasks, where the ability to discern nuanced linguistic cues is paramount. In a comparative evaluation of LSTM and BERT models on cyberbullying text analysis, empirical evidence consistently points towards BERT's superiority in accuracy. Several studies and benchmarking evaluations have demonstrated BERT's ability to outperform LSTM model and other traditional and unique NLP models across a multitude of tasks, including sentiment analysis, text classification, and entity recognition. Its robust performance, coupled with its ability to discern subtle linguistic nuances, positions BERT as the model of choice for cyberbullying detection on Twitter.

One key factor contributing to BERT's superior accuracy is its ability to capture contextual information from both left and right contexts, thereby mitigating the impact of ambiguous or context-dependent language. By considering bidirectional dependencies between words, BERT effectively discerns linguistic nuances and identifies instances of cyberbullying with unprecedented accuracy. Furthermore, BERT's pretrained representations encode rich contextual information, enabling the model to generalize effectively to diverse linguistic styles and nuances present in Twitter data. As a result, BERT exhibits superior accuracy compared to LSTM and other traditional NLP models in cyberbullying text analysis tasks.

Despite BERT's remarkable performance, it is essential to acknowledge potential challenges and limitations associated with its adoption in cyberbullying detection. BERT's computational complexity and resource-intensive nature may pose challenges in deployment, especially in real-time or resource-constrained environments. Additionally, fine-tuning BERT models for specific cyberbullying detection tasks may require large, annotated datasets and substantial computational resources. Moreover, ethical consideration regarding privacy and fairness should be carefully shown to make sure responsible and equitable deployment of BERT-based cyberbullying detection systems.

VII. REFERENCE

[1] Pais, S., Cordeiro, J., & Jamil, M. L. (2022). NLP-based platform as a service: a brief review. Journal of Big Data, 9(1), 1-26.

[2] Yu, J., de Antonio, A., & Villalba-Mora, E. (2022). Deep learning (CNN, RNN) applications for smart homes: a systematic review. Computers, 11(2), 26.

[3] Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., ... & Yuan, L. (2022). Bevt: Bert pretraining of video transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14733-14743).

[4] Cahuantzi, R., Chen, X., & Güttel, S. (2023, July). A comparison of LSTM and GRU networks for learning symbolic sequences. In Science and Information Conference (pp. 771-785). Cham: Springer Nature Switzerland.

[5] Gada, M., Damania, K., & Sankhe, S. (2021, January). Cyberbullying Detection using LSTM-CNN architecture and its applications. In 2021 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.

[6] Iwendi, C., Srivastava, G., Khan, S., & Maddikunta, P. K. R. (2023). Cyberbullying detection solutions based on deep learning architectures. Multimedia Systems, 29(3), 1839-1852..

[7] Ghosh, S., Chaki, A., & Kudeshia, A. (2021, April). Cyberbully detection using 1d-cnn and lstm. In Proceedings of International Conference on Communication, Circuits, and Systems: IC3S 2020 (pp. 295-301). Singapore: Springer Singapore.

[8] Al-Hashedi, M., Soon, L. K., & Goh, H. N. (2019, November). Cyberbullying detection using deep learning and word embeddings: An empirical study. In Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems (pp. 17-21).