DEEP LEARNING ALGORITHM FOR DETECTION OF CYBER BULLYING ACROSS SOCIAL MEDIA PLATOFORMS

Tahira Yousuf[1], Ms. Sara Fatima[2]
[1]PG Scholar, Department of CSE, Shadan Women's College of Engineering and Technology, Hyderabad, Tahira.yousuf271@gmail.com
[2]Asst Professor, Department of CSE, Shadan Women's College of Engineering and Technology,

ABSTRACT

Social networking and communication have been accelerated by information and communication technologies, yet cyberbullying presents serious difficulties. The current laborious and ineffective methods for reporting and stopping cyberbullying are user-dependent. For the automatic detection of cyberbullying, traditional machine learning and transfer learning techniques were investigated. The study made use of a structured annotation procedure and a large dataset. The Conventional Machine Learning method used textual, sentiment and emotional, static and contextual word embeddings, term lists, psycholinguistics, and toxicity characteristics. The use of toxicity features for the detection of cyberbullying was first presented in this study. Word Convolutional Neural Network (Word CNN) contextual embeddings performed similarly; embeddings were selected based on their higher F-measure. When given separately, toxicological features, embeddings, and textual features provide new standards. In terms of managing high-dimensionality features and training time, this performed better than Linear SVC. In contrast to the basis models, Transfer Learning achieved a faster training computation by fine-tuning using Word CNN. Additionally, Flask Web was used to detect cyberbullying, with a 97.06% accuracy rate. For privacy reasons, the name of the particular dataset was not mentioned.

1. INTRODUCTION

With their subtle evolution over time, information and communication technologies (ICT) have become an essential aspect of everyone's life and have sparked online connection between individuals. The increased usage of online platforms has made communication as simple as clicking a button, which has helped social networking to flourish. When people readily abuse technology advancements through abusive behaviors like cyberbullying, ICT dominance has a negative side. The extended version of traditional or direct bullying over electronic platforms is known as cyberbullying. Because social media serves as a virtual platform for bullying and conceals the identity of the perpetrator, identifying cyberbullying is a difficult task that must be done to safeguard online communities. Due to its ease of anonymity, cyberbullying occurrences rise in tandem with increased Internet usage. The goal of this project is to apply the state-of-the-art in NLP and Deep Learning to construct classification models that can accurately identify instances of cyberbullying and non-cyberbullying from disorderly messages. This work combines word CNN model building, feature engineering, and text pre-processing.

OBJECTIVE

Investigate and implement state-of-the-art NLP and Deep Learning techniques to enhance the detection of cyberbullying on online platforms. Examine the inclusion of unique variables, such toxicity indicators, to conventional textual and sentiment data in order to improve the accuracy of cyberbullying detection algorithms.

Create powerful classification models using Word CNN for contextual embeddings that perform better than conventional machine learning methods. As a real-world application of the developed models, incorporate cyberbullying detection into a Flask online platform to attain high accuracy in real-time identification and prevention.

2.1 PROBLEM STATEMENT

On online platforms, cyberbullying has grown to be a major problem that has an impact on users' mental health and general digital wellbeing. The subtle and context-dependent character of hazardous content may be missed by traditional detection methods, which frequently just use sentiment analysis or simple textual analysis. More precise and reliable models that can quickly detect and stop cyberbullying are becoming more and more necessary. In order to improve contextual understanding and classification performance, this project will explore and use cutting-edge Natural Language Processing (NLP) and Deep Learning approaches, particularly Word-level Convolutional Neural Networks (Word CNN). The approach is intended to greatly increase detection accuracy by combining special factors, including toxicity signs, with standard textual and sentiment features.

2.2 EXISTING SYSTEM

With the increasing difficulty of training machine learning (ML) classifiers due to the predominance of

Deep Neural Networks (DNN) and huge datasets, the current approach concentrates on resolving the resource-intensive aspect of this process. The goal of Feature Density (FD) analysis is to optimize the training process by estimating the performance of ML classifiers prior to training.

The study highlights how resource-intensive training has an adverse effect on the environment, particularly in light of the rising CO2 emissions linked to large-scale machine learning models. With a focus on dialog classification, including the detection of cyberbullying, the project attempts to reduce the demands on heavy computer resources and improve efficiency in Natural Language Processing.

Disadvantage of Existing System

- primarily on a density-based criteria, which can leave out nuanced linguistic components crucial for identifying cyberbullying.
- reducing the classifier's ability to recognize complex linguistic patterns and minute indications of cyberbullying.
- Its inability to include complex features and embeddings.

2.3 PROPOSED SYSTEM

The suggested system's automatic detection approach addresses the problem of cyberbullying in social networks. Using a large dataset and structured annotations, the system combines textual, sentimental, emotional, static, and contextual data. This approach is unique in that it enhances the detection of cyberbullying by incorporating toxicity factors.

The system outperforms Linear SVC in terms of handling high-dimensionality features and training time. By optimizing WORD CNN, Transfer Learning improves performance and expedites training computations in comparison to base models. Additionally, a Flask web application ensures a 97.06% accuracy rate in real usage.

Advantages of Proposed System

- It can identify patterns at different degrees of abstraction by detecting intricate semantic relationships in text, which is helpful for identifying a range of cyberbullying phrases.
- Using pre-trained models and expediting training can improve cyberbullying detection without sacrificing performance.

2. RELATED WORKS

A number of techniques provide models for detecting cyberbullying. In the study of [8], the model is proposed to offer a dual definition of cyberbullying by employing a dishonest approach to deal with giving the arrangement with less accuracy and a unique CNN notion for content analysis. The gathered data has been shown to offer better classification and precision when compared to other studies.

[9] published a thorough review of 186 entries from online data repositories. Ten evaluations of the literature have been selected for this article in order to evaluate and discuss the evidence pertaining to machine learning's efficacy in stopping cyberbullying. Most models use content-based factors to forecast cyberbullying. According to research by [10], fuzzy logic was used to develop a method for detecting cyberbullying. This method involves continuously monitoring the two users' conversation and identifying the emotional content of each message. Each user's behavior is categorized as either decent or bullying based on their emotions. If the quantity of bullying that is witnessed beyond a preset threshold, the user's account is immediately closed and reported. They came to the conclusion that it might be a useful tool for preventing online harassment if used in conjunction with social networking sites. The developed method can potentially be applied to human behavior research and monitoring.

3. METHODOLOGY OF PROJECT

In order to combat the growing issue of cyberbullying in online contexts, this research uses cutting-edge Natural Language Processing (NLP) and Deep Learning algorithms. The project intends to build robust categorization models by thoroughly examining a number of characteristics, including indications of toxicity. These algorithms are intended to detect instances of cyberbullying in textual online communication more accurately than traditional machine learning techniques. Word CNN serves as their foundation for contextual embeddings. An example of how these enhancements are implemented is the Flask online platform for real-time cyberbullying detection, which strives for a high degree of identification and preventive accuracy.

MODULE DESCRIPTION:

1) Dataset:

We obtained a dataset for the purpose of detecting cyberbullying during the project's first phase. The collection, which comes from "cyberbullying_tweets.csv," is made up of different text entries i.e., tweets. Included in the dataset are 47,692 information that have been classified as either "not_cyberbullying," "gender," "religion," "other_cyberbullying," "age," or "ethnicity."

2) Importing the Necessary Libraries:

We decided to programme in Python and loaded the necessary project libraries. Key libraries include PIL for turning photos into arrays, scikit-learn for splitting the data into training and testing sets, and other common libraries like pandas, numpy, matplotlib, and TensorFlow. Keras is used to build the primary model.

3) Data Pre-processing:

We used pandas to read the CSV file, info() to do a first data inspection, and handling any missing values. The 'cyberbullying_type' labels were then encoded using Label Encoding, which changed the textual data. Next, we divided the dataset into sets for testing, validation, and training.

4) Model Creation for Word CNN:

We use a word Convolutional Neural Network (CNN) as they have proven to be successful at document classification problems. A conservative word CNN configuration is used with 128 filters (parallel fields for processing words) and a kernel size of 5 with a rectified linear ('relu') activation function. This is followed by a pooling layer that reduces the output of the convolutional layer.

We can see that the Embedding layer expects documents with words as input and encodes each word in the document as a 11element vector.

We use a categorical cross entropy loss function because the problem we are learning is a categorical classification problem. The efficient Adam implementation of stochastic gradient descent is used and we keep track of accuracy in addition to loss during training. The model is trained for 35 epochs, or 64 passes through the training data. The last dense layer outputs 6 nodes as either "not_cyberbullying," "gender," "religion," "other_cyberbullying," "age," or "ethnicity". This layer uses the softmax activation function which gives probability value and predicts which of the 6 options has the highest probability.

5) Training and Evaluation:

Using the fit function, we trained the Word CNN model by setting hyperparameters such batch size and epochs. An average of 97.06% was achieved in training, while an average of 99.92% was achieved in validation. We then assessed the model using the test set, and 97.7% accuracy was obtained.

6) Saving the Trained Model:

We proceeded to preserve the model for UI final detection/test use. Using the Tensorflow/ Keras library, the model was stored in the Hierarchical Data Format (HDF5). Furthermore, the pickle library was used to serialise the tokenizer that was used for text preprocessing, which was then saved as "tokenizer.pickle".

4. ALGORITHM USED IN PROJECT

The proposed approach utilizes Word CNN's capabilities to detect cyberbullying. Transfer Learning adapts the WORD CNN to the distinct features of cyberbullying in the dataset. Compared to starting from scratch, this process speeds up training calculation and enhances the model's ability to recognize subtle patterns.

Crucially, the WORD CNN's contextual embeddings function similarly when the F-measures are bigger. By extending the conventional use of embeddings and offering a novel interpretation of toxicity features, our approach contributes to a more comprehensive cyberbullying detection model. The system's usage of WORD CNN not only increases accuracy but also demonstrates a forward-thinking strategy for addressing the evolving challenges of online social interactions.

5. DATA FLOW DIAGRAM

Level 0

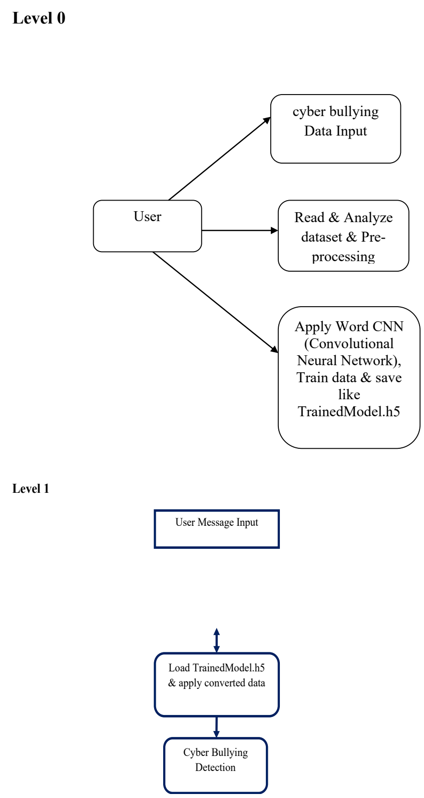


Fig: 6 Flow Diagram

6. SYSTEM ARCHITECTURE

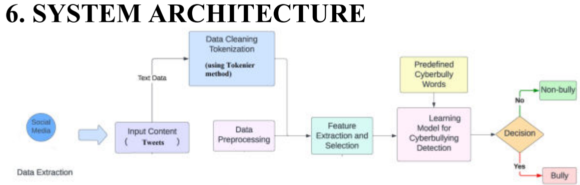
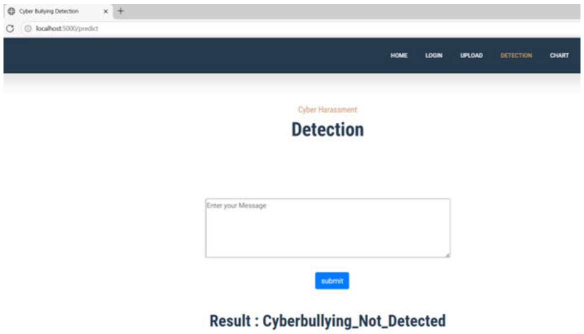
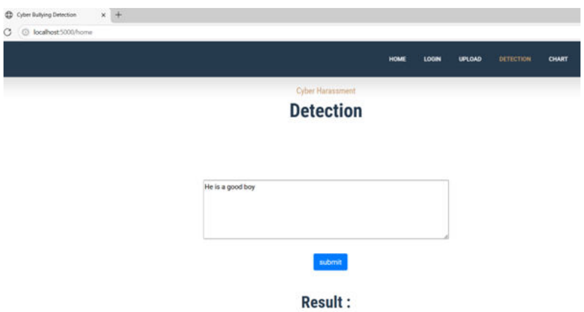


Fig: 7 SYSTEM ARCHITECTURE OF PROJECT

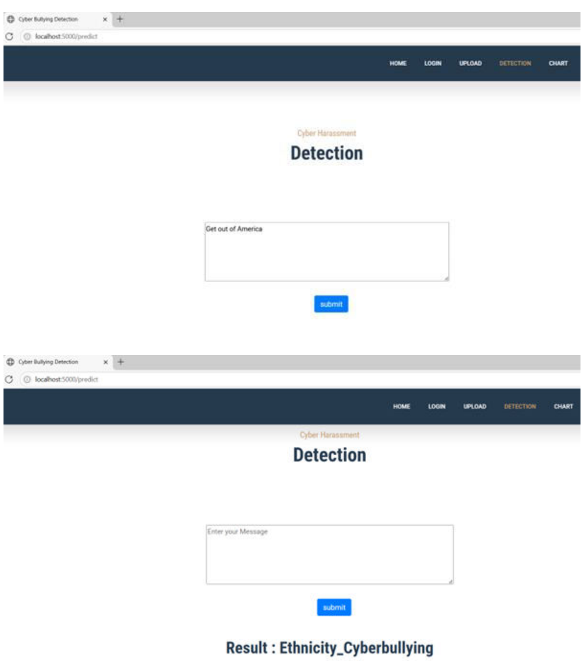
7. RESULTS



Input Sample



Result Sample



Detection Sample

8. FUTURE ENHANCEMENT

Even though the cyberbullying corpus includes input from several roles within cyberbullying episodes, our approach is restricted to binary text categorization and does not assist us in identifying the poster of the cyberbullying post. The participation roles of harassers, victims, onlookers, and non-bullies can all be categorized using this approach. The research might be expanded to take into consideration the relationships between postings in order to capture user engagement throughout cyberbullying incidents, as the conversation was handled as a single post.

9. CONCLUSION

In summary, the unexpected increase in cyberbullying brought about by technological innovation has brought attention to the urgent need for effective preventive measures. Since automated detection techniques have the potential to have serious and widespread effects on Internet users, they must be developed and implemented. This is a prophylactic measure that also significantly lowers the frequency of instances of cyberbullying. While previous methods for categorizing cyberbullying have primarily relied on textual features, this study has adopted a more comprehensive approach by examining a wide range of feature categories. By looking at textual features, sentiment and emotional features, embeddings, psycholinguistic features, word list characteristics, and toxicity factors, we have expanded the pool of potential indicators for cyberbullying identification.

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

[16]. D. Noever, ''Machine learning suites for online toxicity detection,'' 2018, arXiv:1810.01869.