International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

DEEP LEARNING ALGORITHMS FOR CYBER-BULLING DETECTION IN SOCIAL MEDIA PLATFORMS

K. VIJAYA LAXMI [1], BADOLA ARYAN [2], POTHURAJU VENKATESH [3], BADUGULA AJAY KUMAR [4], JATAVATH JAIDEEP NAIK [5]

[1]Assistant Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

[2, 3, 4, 5] B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

ABSTRACT— Information and Communication Technologies have propelled social networking and communication, but cyber bullying poses significant challenges. Existing user-dependent mechanisms for reporting and blocking cyber bullying are manual and inefficient. Conventional Machine Learning and Transfer Learning approaches were explored for automatic cyber bullying detection. The study utilized a comprehensive dataset and structured annotation process. Textual, sentiment and emotional, static and contextual word embeddings, psycholinguistics, term lists, and toxicity features were employed in the Conventional Machine Learning approach. This research introduced the use of toxicity features for cyber bullying detection. Contextual embeddings of word Convolutional Neural Network (Word CNN) demonstrated comparable performance, with embeddings chosen for its higher F-measure. Textual features, embeddings, and toxicity features set new benchmarks when fed individually. This outperformed Linear SVC in terms of training time and handling high-dimensionality features. Transfer Learning utilized Word CNN for fine-tuning, achieving a faster training computation compared to the base models. Additionally, cyber bullying detection through Flask web was implemented, yielding an accuracy of

97.06%. The reference to the specific dataset name was omitted for privacy.

Index Terms – Cyber Harassment Prediction, Word CNN, Social Media, Natural Language Processing, Deep Learning, Text Classification, Cyberbullying Detection, Offensive Language Detection.

I. INTRODUCTION

Information and Communication Technologies (ICT) have become an integral part of everyone's life, evolving imperceptibly with time, catalyzing online communication between people. Communication has been just one button click with the widespread use of the online platform, facilitating the growth of social networking. ICT dominance has a dark side when people easily misuse technological advancement with abusive behaviors such as cyberbullying. Cyberbullying is the expanded form of direct or traditional bullying through electronic platforms. Social media becomes the virtual medium for bullying, shielding the bully's identity, making detecting cyberbullying a complex and challenging mission to protect online communities. Cyberbullying cases increase with volumized Internet usage because it can be easily committed anonymously, leading to a grave public health concern that brings many negative impacts, such as mental, psychological, and social problems. While cyberbullying victims tend to suffer from mental health problems such as depression, anxiety, loneliness, and anhedonia, some are reported to be committing self-injurious behavior and suicidal ideation.

The expected outcome of this research is the development of classification models that can effectively detect cyberbullying and non-cyberbullying events from unruly posts by applying the knowledge of state of the art in NLP and Deep Learning. This work incorporates text pre-processing, feature engineering, model development using word CNN.

II. RELATED WORK

A. Cyber Bullying Detection Using Machine Learning

Cyber bullying has evolved as a severe problem hurting children, teenagers, and young adults as a result of the increasing use of social media. Automatic detection of bullying communications in social media is now possible, thanks to machine learning techniques, which could aid in the creation of a healthy and safe social media environment. One major issue in this important research area is robust and discriminative numerical

representation learning of text messages. To address this challenge, we offer a new representation learning method in this study. The Semantic-Enhanced Marginalized Denoising Auto-Encoder (SMSDA) is a semantic enhancement of the popular deep learning model stacked denoising Auto-Encoder. The semantic extension is made up of semantic dropout noise and sparsity constraints, with the semantic dropout noise being the most important.

B. Cyber Bullying Detection for Twitter Using ML Classification Algorithms

Social networking platforms have given us incalculable opportunities than ever before, and its benefits are undeniable. Despite benefits, people may be humiliated, insulted, bullied, and harassed by anonymous users, strangers, or peers. Cyberbullying refers to the use of technology to humiliate and slander other people. It takes form of hate messages sent through social media and emails. With the exponential increase of social media users, cyberbullying has been emerged as a form of bullying through electronic messages. We have tried to propose a possible solution for the above problem, our project aims to detect cyberbullying in tweets using ML Classification algorithms like Naïve Bayes, KNN, Decision Tree, Random Forest, Support Vector etc. and also we will apply the NLTK (Natural language toolkit) which consist of bigram, trigram, n-gram and unigram on Naïve Bayes to check its accuracy. Finally, we will compare the results of proposed and baseline features with other machine learning algorithms. Findings of the comparison indicate the significance of the proposed features in cyberbullying detection. The study reviewed and identified Naïve bayes N-gram gives the best accuracy and also the system is able to identify the bullied and non-bullied statements.

C. Detecting A Twitter Cyberbullying Using Machine Learning

Social media is a platform where many young people are getting bullied. As social networking sites are increasing, cyberbullying is increasing day by day. To identify word similarities in the tweets made by bullies and make use of machine learning and can develop an ML model automatically detect social media bullying actions. However, many social media bullying detection techniques have been implemented, but many of them were textual based. The goal of this paper is to show the implementation of software that will detect bullied tweets, posts, etc. A machine learning

model is proposed to detect and prevent bullying on Twitter. Two classifiers i.e. SVM and Naïve Bayes are used for training and testing the social media bullying content. Both Naive Bayes and SVM (Support Vector Machine) were able to detect the true positives with 71.25% and 52.70% accuracy respectively. But SVM outperforms Naive Bayes of similar work on the same dataset. Also, Twitter API is used to fetch tweets and tweets are passed to the model to detect whether the tweets are bullying or not.

D. Brute-Force Sentence Pattern Extortion from Harmful Messages for Cyberbullying Detection

Cyberbullying, or humiliating people using the Internet, has existed almost since the beginning of Internet communication. The relatively recent introduction of smartphones and tablet computers has caused cyberbullying to evolve into a serious social problem. In Japan, members of a parent-teacher association (PTA) attempted to address the problem by scanning the Internet for cyberbullying entries. To help these PTA members and other interested parties confront this difficult task we propose a novel method for automatic detection of malicious Internet content. This method is based on a combinatorial approach resembling brute-force search algorithms, but applied in language classification. The method extracts sophisticated patterns from sentences and uses them in classification. The experiments performed on actual cyberbullying data reveal an advantage of our method vis-à-vis previous methods. Next, we implemented the method into an application for Android smartphones to automatically detect possible harmful content in messages. The method performed well in the Android environment, but still needs to be optimized for time efficiency in order to be used in practice.

III. PROPOSED METHODOLOGY

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.

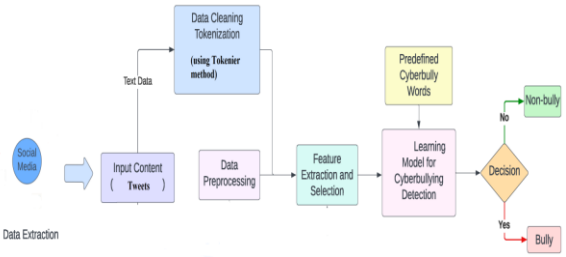


Fig. 1: System Overview

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".

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed cyber harassment prediction system using a Word CNN model was evaluated using performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis. Experimental results showed that the model effectively identified and classified harassing, abusive, and offensive content from social media text data with high accuracy. The Word CNN architecture successfully captured contextual and semantic relationships between words, leading to improved detection performance compared to traditional machine learning methods. The analysis demonstrates that the proposed approach provides reliable and efficient real-time monitoring of harmful online content, contributing to safer social media environments.



Fig. 2: Home Page

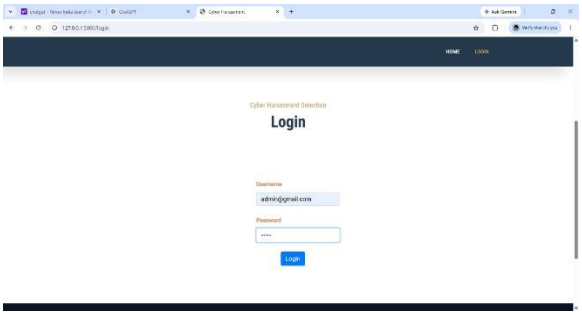


Fig. 3: Login

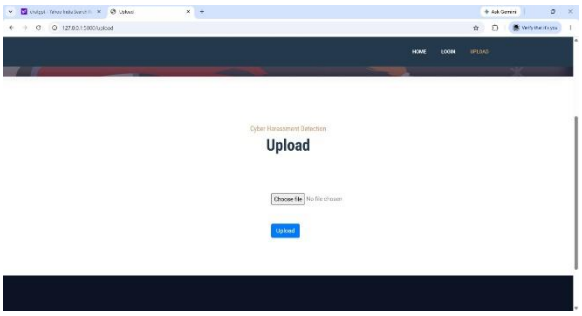


Fig. 4: Upload

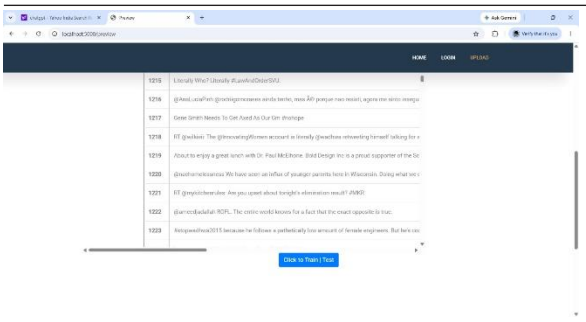


Fig. 5: Train and Test

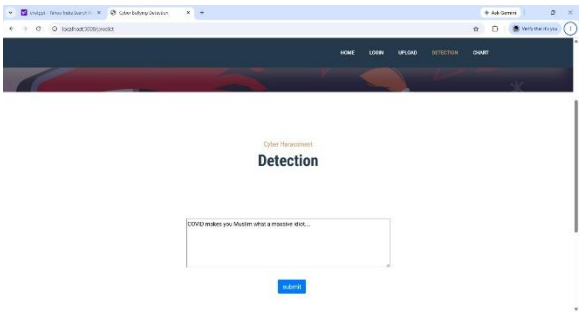


Fig. 6: Enter Text

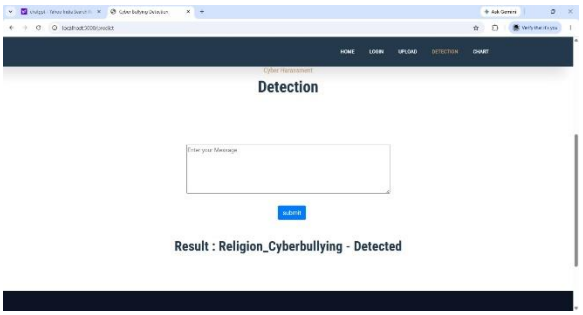


Fig. 7: Detection

V. 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.

AUTHORS Profile



Ms. K. VIJAYA LAXMI received her B. Tech degree in Computer Science & Engineering from JNTUH, Hyderabad in 2015 and her M. Tech degree in Computer Science from JNTUH, Hyderabad in 2017. She is currently pursuing her Ph.D. degree in Computer Science & Engineering from Guru Nanak University, Hyderabad. She has been dedicated to the teaching field for the last 7 years. Her research area is Deep Learning. At present, she is working as an Assistant Professor in the CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India.



Mr. BADOLA ARYAN pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. POTHURAJU VENKATESH pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. BADUGULA AJAY KUMAR pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.

Mr. JATAVATH JAIDEEP NAIK pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.