BULLY NET: UNMASKING CYBER BULLIES ON SOCIAL NETWORK

Seenaiha[1], Stephen Mutluri[2], Vishnu Vardhan Vanka[2], VVS. Srivathsa[2]

[2]UG Scholar, [1,2]Department of Computer Science and Engineering

[1,2]Kommuri Pratap Reddy Institute of Technology, Ghatkesar, Hyderabad, Telangana.

ABSTRACT

Cyberbullying, a serious social issue, involves using the internet and electronic devices to harm or harass others, especially with the rise of social media platforms like Facebook and Twitter. Detecting cyberbullying is complex because traditional search engines overlook semantic information, focusing only on user keywords. Victims, particularly adolescents, suffer from depression, sleeplessness, low self-esteem, and even suicidal thoughts. While efforts have been made to combat traditional bullying, similar attention is needed for cyberbullying. Traditional detection methods, such as user guidelines, manual editing, and profane synonym lists, are time-consuming and ineffective for social media's scale. Thus, an automated and accurate learning framework is essential. Our research targets textual cyberbullying detection, the most common form of communication on social media. Social media content is often short, noisy, and unstructured, with spelling errors and symbols that hinder traditional machine learning methods. We propose a Char-CNN (Character-level Convolutional Neural Network) model to identify cyberbullying in social media text. By using characters as the smallest unit of learning, the model can effectively handle spelling errors and intentional obfuscation, providing a robust solution for detecting cyberbullying and ensuring safer online environments.

Keywords: Convolutional Neural Network, TFIDF, social media, Cyber bullying, Natural Language processing, spellings.

1. INTRODUCTION

Cyberbullying is an increasingly important and serious social problem, which can negatively affect individuals. It is defined as the phenomena of using the internet, cell phones and other electronic devices to willfully hurt or harass others. Due to the recent popularity and growth of social media platforms such as Facebook and Twitter, cyberbullying is becoming more and more prevalent. Many applications of the World Wide Web need to discover the envisioned meaning of certain textual resources (e.g., data to be annotated, or keywords to be searched) in order to semantically describe the result causing the effects, such as the abusive words usage causes to create the impact of cyberbullying. However, this cyberbullying detection is more complicated because current search engine focusses only on retrieving the results containing the user keywords, and lots of data that may carry the desired semantic information remains overdue. The cyber cyberbullying detection is advanced topic in Artificial Intelligence research and related fields, which is a major problem not only in NLP but in the Semantic Web services as well. Disambiguation methods mean to get the most suitable sense of an ambiguous word according to the context. Cyberbullying is bullying that takes place over digital devices such as cell phones, computers, and tablets [1]. Cyberbullying can be achieved in various ways, such as sending a message containing abusive or offensive content to a victim, and some labeled posts are shown in Table 1. In a 2018 statistical report, during the 2015-16

school year, approximately 12% of public schools reported that students had experienced cyberbullying on and off campus at least once a week, and 7% of public schools reported that the school environment was affected by cyberbullying [2]. It can create negative online reputations for victims, which will impact college admissions, employment, and other areas of life, and can result in even more serious and permanent consequences such as self-harm and suicide [3]. Cyberbullying events are hard to recognize. The major problem in cyberbullying detection is the lack of identifiable parameters and clearly quantifiable standards and definitions that can classify posts as bullying [4]. As people spend increasingly more time on social networks, cyberbullying has become a social problem that needs to be solved, and it is very necessary to detect the occurrence of cyberbullying through an automated method. Our research focuses on textual cyberbullying detection because text is the most common form of social media. In text-based cyberbullying detection, capturing knowledge from text messages is the most critical part, but it is still a challenge. The first challenge that cannot be ignored is dealing with unstructured data. The content information in social media is short, noisy, and unstructured with incorrect spellings and symbols [5] such as the instances in Table 1. Social media users intentionally obfuscate the words or phrases in the sentence to evade manual and automatic detection as in R3. These extra words will expand the size of the vocabulary and influence the performance of the algorithm. Emojis made up of symbols such as :) in R4, which definitely convey emotional features, are always hard to distinguish from noise.

Table 1: Some instances in dataset.

R1	Sassy. More like trashy
R2	I HATE KAT SO MUCH
R3	Kat, a massive c*nt
R4	Shut up Nikki… That is all :)

Another key challenge in cyberbullying research is the availability of suitable data, which is necessary for developing models that can classify cyberbullying. There are some datasets have been publicly available for this specific task such as the training set provided in CAW 2.0 Workshop and the Twitter Bullying Traces dataset [6]. Since cyberbullying detection has been fully illustrated as a natural language processing task, various classifiers have been masterly improved to accomplish this task, including the Naive Bayes [7], the C4.5 decision tree [8], random forests [9], SVMs with different kernels, and neural networks classifiers [6]. A variety of feature selection methods have also been carefully designed to improve the classification accuracy.9-13 However, previous data-based works have relied almost entirely on vocabulary knowledge, and so, the challenges that are posed by unstructured data still exist.

2. LITERATURE SURVEY

Akhter, et al. [1] proposed a robust hybrid ML model for cyberbullying detection in the Bengali language on social media. The Bengali bullying proposal involved effective text preprocessing to convert Bengali text data into a useful text format, feature extraction using the Tfid Vectorizer (TFID) to obtain beneficial information from text data, and resampling by Instance Hardness Threshold (IHT) procedure to balance the dataset to avoid overfitting or underfitting problems. In their experiment, they

used the publicly available Bangla text dataset (44,001 comments) and achieved the highest performance ever published on it.

Iwendi, et al. [2] proposed different models based on deep learning algorithms to make an impact on the detection of cyberbullying. These detection mechanisms resulted in efficient identification of incidents while others had limitations of standard identification versions. Their paper performed an empirical analysis to determine the effectiveness and performance of deep learning algorithms in detecting insults in Social Commentary. They used four deep learning models for experimental results, namely Ali, et al. [3] Launched an annotated large-scale dataset with approximately 14,000 English tweets, which was used in various works for detecting offensive posts in social media, predicting their type and target. To solve the classification problems associated with the OLID dataset, nine supervised models were developed for each task of the dataset. Sultan, et al.[4] Determined that cyberbullying frequently leads to severe emotional and physical suffering, especially in women and young children. In certain instances, it was reported that sufferers attempted suicide. The bully was occasionally attempt to destroy any proof they believe to be on their side. Even if the victim got the evidence, it would still be a long time before they got justice at that point Yi, et al. [5] proposed evidence-based criteria for a set of best practices to create session-based cyberbullying datasets. In addition, they performed benchmark experiments comparing the performance of state-of-the-art session-based cyberbullying detection models as well as large pre-trained language models across two different datasets. Through their review, they also put forth a set of open challenges as future research directions.

Mahajan, et al. [6] proposed a model that utilized Bi-directional Long Short-Term Memory (BiLSTM), Bi-directional Gated Recurrent Unit (Bi-GRU), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) techniques to enhance the overall performance. They first preprocessed the multilingual data streams followed by adoption of Global vectors for word Representation (GloVe) embeddings to convert words to a vector representation in parallel enabling the data streams for binary classification task. In order to construct an architecture for the detection of hate speech and cyberbully, they introduced a heterogeneous fusion of multiple effective models in a unique approach such that CNN-LSTM utilized a stacking approach with stochastic gradient descent to achieve optimal weights, whereas all the base learners used a bagging ensemble approach with cross-validation to reach optimal weights

Mkwananzi, et al. [7] Launched research aim was achieved through a mixed methods research design, containing qualitative and quantitative elements. The drivers of cyberbullying were identified through a literature review. These included age, gender, family structure, parental education, race, technology, anonymity, academic achievement, and awareness of cyber safety. The support vector machines and naïve Bayes models were used to classify the text dataset (Formspring.me dataset), with a 72.81% and a 99.87% classification accuracy, respectively. Murshed, et al. [8] proposed a new CB detection model named FAEO-ECNN for detecting and classifying cyberbullying on social media platforms. The proposed approach integrated Fuzzy Adaptive Equilibrium Optimization (FAEO) clustering-based topic modelling and Extended Convolutional Neural Network (ECNN) to enhance the accuracy of CB detection process. Initially, pre-processing was performed in order to cleanse the dataset. Next, the features were extracted using multiple models. Neha, et al. [9] proposed a comparative analysis of various automated cyberbullying detection techniques. Additionally, they recommended a comprehensive strategy for preventing cyberbullying and emphasized the significance of dynamical prediction algorithms in detecting cyberbullying. They also discussed the possible threat of

cyberbullying in the emerging metaverse concept. The chapter highlighted the technical challenges associated with cyberbullying detection and identified areas of research that required attention to improve cyberbullying prevention in the future. By increasing awareness and promoting responsible digital citizenship, they aimed to combat cyberbullying and promote a more positive and inclusive online community.

Ahmed, et al. [10] proposed that recently, Bangla text had been used much more often on social media. People communicated with others on social media through messages and comments. So, bullies used social media as a rich environment to bully others, especially on political issues. Fights over Cyberbullying on political and social media posts were common at that time. Most of the time, it did a lot of damage Al-Ajlan, et al. [11] proposed an algorithm called Cyberbullying Detection Based on Deep Learning (CDDL) to bridge this gap. It eliminated the need for feature engineering and generated better predictions than traditional approaches for detecting cyberbullying. This was accomplished by incorporating deep learning—specifically, a convolutional neural network (CNN)—into the detection process. HUANG, et al. [12] Proposed The data from Formspring.me, a question-and-answer formatted website, was labeled by Amazon's Mechanical Turk, a web service. After data preparation and pre-processing, Scikit-learn, a Python library, was used to train a model to classify if cyberbullying was present in the post by recognizing bullying content based on its insult words as features. Four machine learning techniques, namely, logistic regression, decision tree, random forest, and support vector machine, were used to train the model. Fati, et al. [13] The proposed cyberbullying detection methods were evaluated using benchmark experimental datasets and well-known evaluation measures. Finally, the results demonstrated the superiority of the attention-based 1D convolutional long short-term memory (Conv1DLSTM) classifier over the other implemented methods.

3. PROPOSED METHODOLOGY

The research uses various libraries like Tkinter for GUI, Keras for building and training neural network models, NLTK for natural language processing tasks, and Matplotlib for data visualization. The main purpose of the program seems to be cyberbullying detection at both word and character levels using CNNs. Data preprocessing plays a crucial role in the program's workflow. The clean_doc function removes punctuation, stop words, and non-alphabetic characters from text messages, preparing them for tokenization. Additionally, there's an extension_clean_doc function, extending the cleaning process by incorporating WordNet from NLTK to include synonyms of words. This expands the vocabulary and potentially enhances the model's understanding of text context. The program provides functionality for uploading a dataset, cleaning the data, and preparing it for training. It includes functions to clean the text data by removing punctuation, stop words, and non-alphabetic characters. Additionally, there are functions to tokenize the text data, encode it into sequences, and calculate various metrics such as precision, recall, F1-score, and accuracy. It also supports the training of CNN models for cyberbullying detection based on both word-level and character-level representations of the text data. The CNN models are constructed using Keras, and the program includes functions to train these models, save their weights and architectures, and evaluate their performance on test data. Overall, the program aims to provide a user-friendly interface for cyberbullying detection tasks using machine learning techniques.the program serves as a comprehensive tool for cyberbullying detection, offering users a user-friendly interface to upload datasets, preprocess text data, train CNN models, evaluate model

performance, and visualize results. Its modular design, leveraging various libraries and techniques, makes it a valuable asset for researchers and practitioners working in the field of natural language processing and cyberbullying prevention.

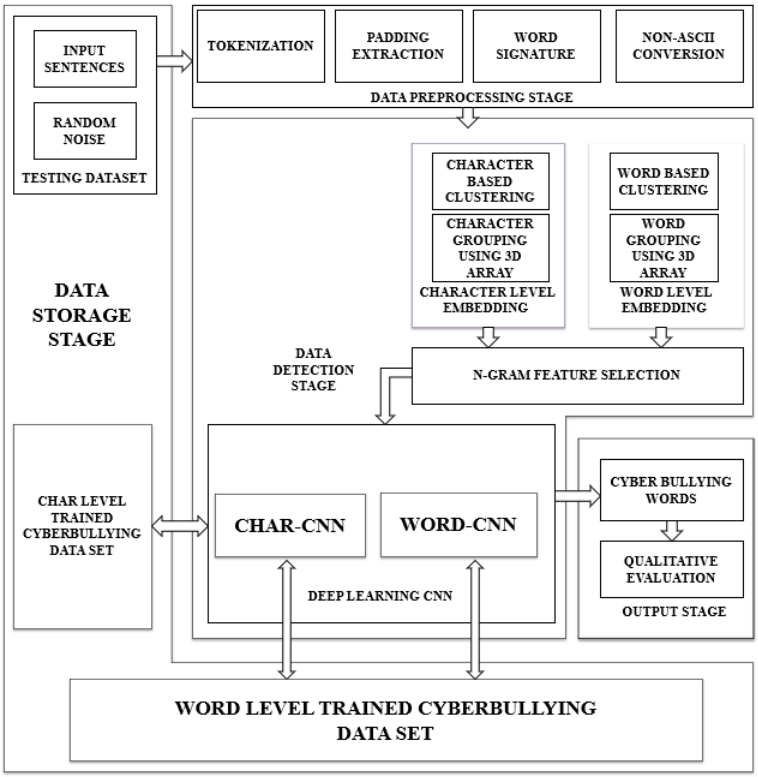

Figure 1: Proposed cyberbullying detection architecture.

Word signature

Unknown word handling module Unknown words are defined as the words which are not in the lexicon or in reference sentences. Since CNN algorithm generate error as it detects unknown word therefore a separate module is required for tag decision for unknown word. In case of cyberbullying scenario, the attackers use the complicated abusive words; they may not be presented in the vocabulary. Thus, out of vocabulary words also considered for cyberbullying detection.

Non-ASCII conversion

Electronic processing of text in any language requires that characters (letters of the alphabet along with special symbols) be represented through unique codes, this is called encoding. Usually, this code will also correspond to the written shape of the letter. A NON-ASCII conversion is basically a number associated with each letter so that computers can distinguish between different letters through their codes.

CNN architecture

This section describes about the implementation details of HCNN based approach for cyberbullying detection with respect to character, and word recognition. Figure 1 represents the seven layered deep

learning network architecture with F number of filers, Kernel size as K; it consists of multiple hidden layers which will allow it to compute much more complex features of the input.

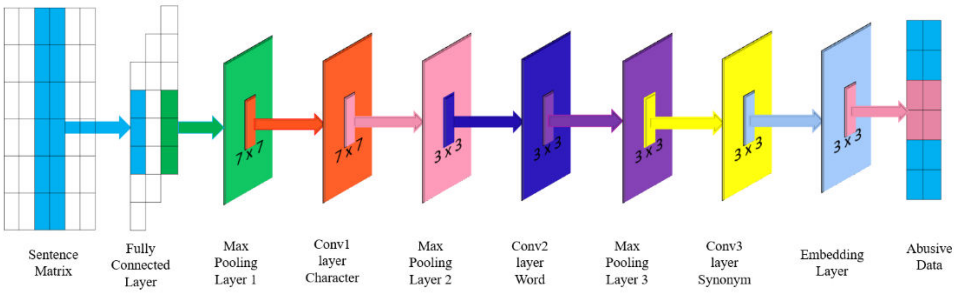

Figure 2: CNN model.

Because each hidden layer computes a non-linear transformation of the previous layer, a deepnetwork can have significantly greater representational power (i.e., can learn significantly more complex functions) than a shallow one. By using a deep network, in the case of text data, one can also start to learn part-whole decompositions. For example, the first layer learns to group together characters in text to detect abusive data. The second and third layers might then group together words to detect cyberbullying content, or perhaps detect simple "parts of objects." An even deeper layer might then group together these word-based synonyms or detects even more complex features.

Initially the features of the sentence will be applied as matrix to the fully connected layer. The fully connected layer extracts the features of input text samples and parallel it will map the different types of features of text. The fully connected layer is used for binary classification of data. In computer vision tasks pooling layers are introduced for invariance of the nodes to rotation, translation, or scaling, modelling that textual data is the same, even occurring on another patch of the data, or in another size. Another handy property is that a variable sized feature map is thus reduced to a fixed size entity. Another handy attribute of Max pooling layer is that the network now can handle differently sized texts because despite the size of the input tensor or the then computed feature map the pooling operation yields only one value. Also pooling layers can make features invariant to translation or scaling because of its independency of the features position in the feature map.

Convolutions in the context of machine learning are filters that are applied to receptive fields of an input tensor, yielding tensors of filter maps representing the activation of a region on the inputs. In this case the input tensors are simply matrices, and the individual feature maps are scalar due to the dot product of the weight vector w and the sentence matrix A. The convolution is applied to a matrix by computing the filter F on windows where each window represents a slice of the matrix as given by a stride width and height. Words in each row are reduced row wise using stride, thus viewing words in 3D minute form. A filter F of height h is applied to a window of words (i representing number of rows and j as number of columns). Rows of the matrix represent a word in NLP applications. Thus, it develops $i: j + h - 1$ scenario in convolution network. Choosing a filter that is applied with a column-wise stride too can be used for convolutions on characters in the word. A convolution is defined by a K kernel function (activation function) that is applied to the patches as a weight matrix and a chosen non-linearity. In the forward pass, when applying convolutional computing, each filter F is slid across the width and height and the dot product is computed between the entries of filter and the corresponding inputs.

The input given is the univariate vector formed from each token of a sentence in the form of a matrix. The filters slide over full rows of the matrix (words) to identify the most important feature in the sentence. The experimentation is done with 16, 32 and 64 number of filters. The "width" of the filters is usually the same as the width of the input matrix. The size of different layers are [64, 7, 3], [64, 7, 3], [64, 3, -1], [64, 3, -1], [64, 3, -1], [64, 3, 3] and [64, 3, -1] respectively. In each layer 64 refers to number of filters F used to build text features and 7 refers to kernel size K of each filter and 3 refers to 3-dimension array of pool size P. As our features will have 3-dimensional data. -1 refers to output layer where we get output data as two-dimensional arrays. Here, max pooling layer 1 and convolutional layer1 will be used to detect the character-based cyberbullying. And max pooling layer 2 and convolutional layer 2 will be used to detect the word-based cyberbullying. And max pooling layer 3 and convolutional layer 3 will be used to detect the synonym-based cyberbullying. Results of the different layers will be combined across the embedding layer and results in the abusive data

4. RESULTS AND DISUSSION

Convolution Neural Networks was designed for image processing but it also giving best performance in Natural Language Processing to detect sentiments from text or cyberbullying. Existing techniques were using words vector to embed or feed data into CNN networks and these networks may not predict correct class due to small spelling mistakes available in train data and sometime some users may give spelling mistakes to avoid detection process. To allow CNN network to predict spelling mistakes or shortcuts data we are building Character Based CNN networks. To design character-based CNN we will split text data into words and then extract characters from each work and build a vector. CNN embedding layer can be created using all characters available in English language and this embedding layer act as vocabulary for CNN. CNN filter all text data based on embedding layer.

Vocabulary example for CNN 'a:0,b:1,c:2,d:3 and goes on for all characters'

If user give input as 'bc' then CNN convert 'b, c' with embedding weight such as '1,3' as b is available at index 1 and c available at index 2. Similarly, CNN will build model by scanning embedding vocabulary.

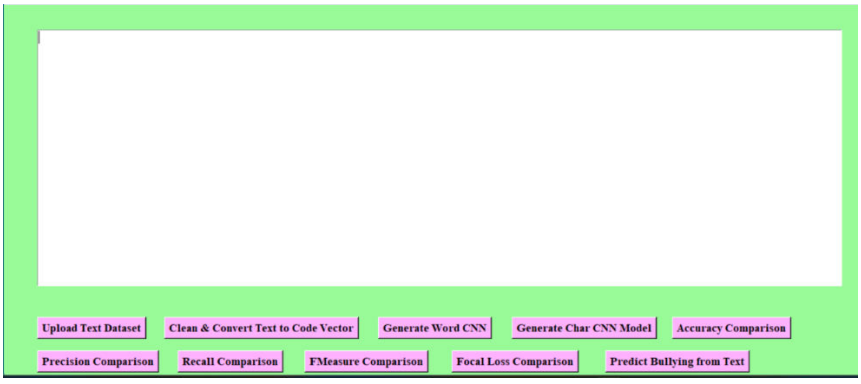

Figure 3: Sample UI Used for Cyber Bullying detection

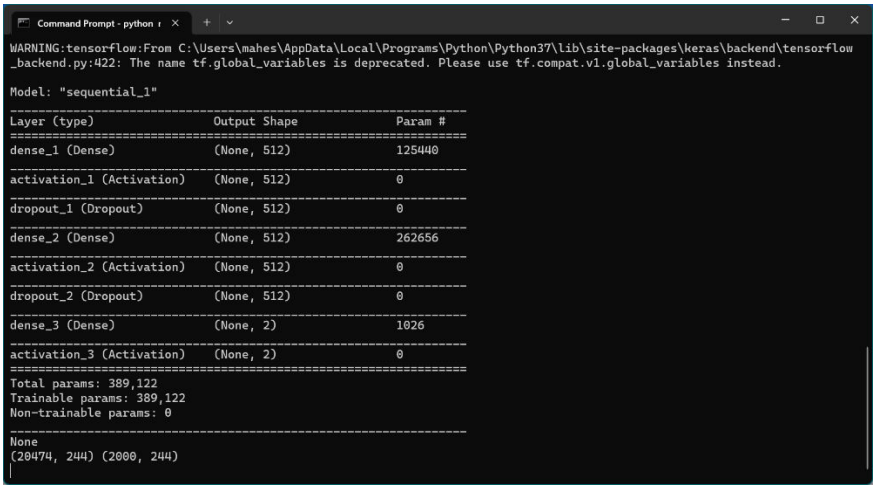

Figure 4: CNN Model Summary.

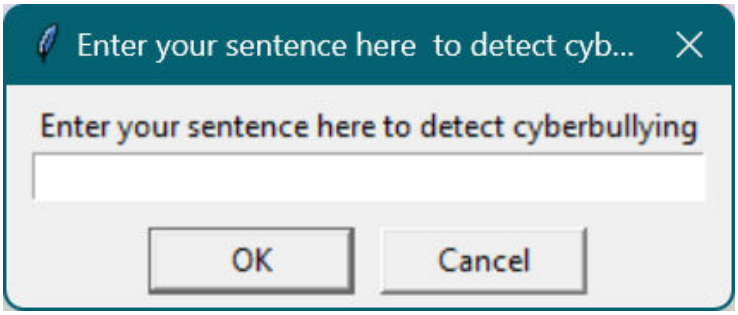

Figure 5: Presents the upload of input sentence.

In above screen I entered text message as 'Hi... how are you man' and below is the prediction result.

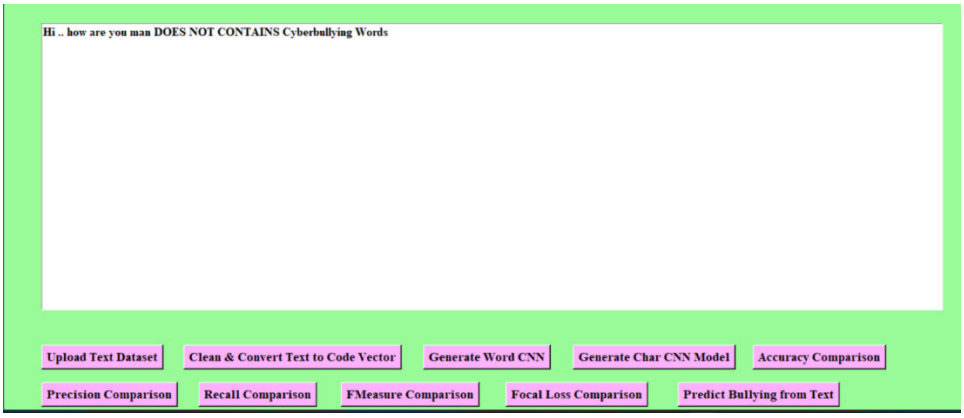

Figure 6: Model predication of Uploaded input sentence.

In above screen model predicted that given text message does not contain any cyber bullying words. Now will test with another sentence.

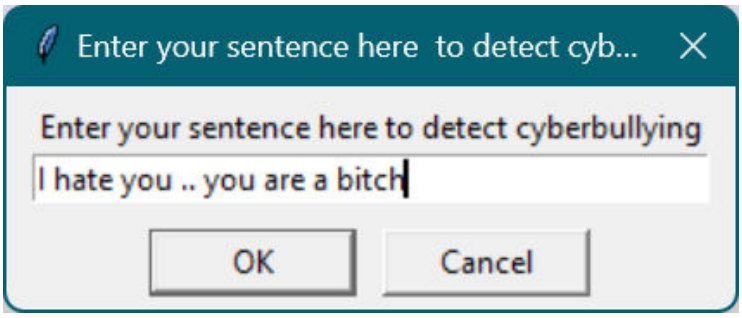

Figure 7: Presents the upload of input sentence for model prediction.

In above screen I entered message as 'I hate you. you are a bitch' and below is the result.

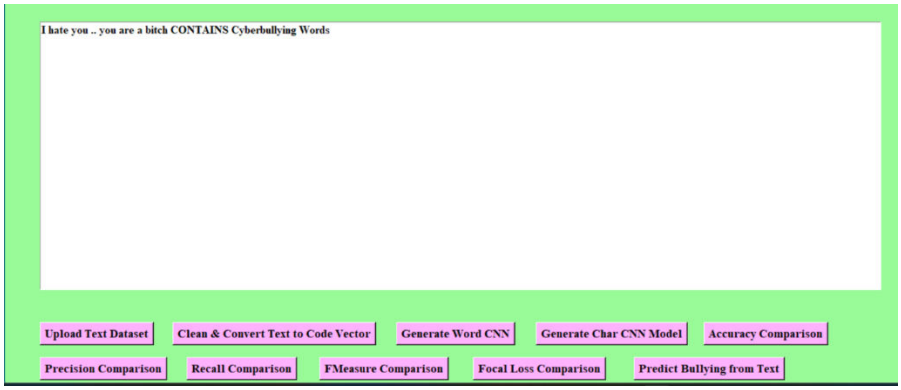

Figure 8: Model predication of Uploaded input sentence.

In above screen we got prediction result as given message contains Cyber Bullying words. Below is an example of predicting bullying with shortcuts.

5. CONCLUSION

The Research Detecting Cyberbullies on social media in the age of Big Data: a machine learning approach has successfully demonstrated the effectiveness of utilizing machine learning techniques to identify and mitigate cyberbullying on social media platforms. By analyzing large volumes of data, the system has shown promising results in accurately detecting instances of cyberbullying, thereby contributing to the creation of safer online environments. Additionally, the project highlights the importance of leveraging big data and machine learning in addressing contemporary social issues such as cyberbullying.

REFERENCES

[1] Akhter, Arnisha, Uzzal Kumar Acharjee, Md Alamin Talukder, Md Manowarul Islam, and Md Ashraf Uddin. "A robust hybrid machine learning model for Bengali cyber bullying detection in social media." *Natural Language Processing Journal* 4 (2023): 100027.

[2] Iwendi, Celestine, Gautam Srivastava, Suleman Khan, and Praveen Kumar Reddy Maddikunta. "Cyberbullying detection solutions based on deep learning architectures." *Multimedia Systems* 29, no. 3 (2023): 1839-1852.

[3] Ali, Mohammad Usmaan, and Raluca Lefticaru. "Detection of cyberbullying on social media platforms using machine learning." In *UK Workshop on Computational Intelligence*, pp. 220-233. Cham: Springer Nature Switzerland, 2023.
[4] Sultan, Tofayet, Nusrat Jahan, Ritu Basak, Mohammed Shaheen Alam Jony, and Rashidul Hasan Nabil. "Machine learning in cyberbullying detection from social-media image or screenshot with optical character recognition." *Int. J. Intell. Syst. Appl* 15 (2023): 1-13.
[5] Yi, Peiling, and Arkaitz Zubiaga. "Session-based cyberbullying detection in social media: A survey." *Online Social Networks and Media* 36 (2023): 100250.
[6] Mahajan, Esshaan, Hemaank Mahajan, and Sanjay Kumar. "EnsMulHateCyb: Multilingual hate speech and cyberbully detection in online social media." *Expert Systems with Applications* 236 (2024): 121228.
[7] Mkwananzi, Nomandla, and Hanlie Smuts. "Guidelines for Detecting Cyberbullying in Social Media Data Through Text Analysis." *International Journal of Social Media and Online Communities (IJSMOC)* 15, no. 1 (2023): 1-13.
[8] Murshed, Belal Abdullah Hezam, Suresha, Jemal Abawajy, Mufeed Ahmed Naji Saif, Hudhaifa Mohammed Abdulwahab, and Fahd A. Ghanem. "FAEO-ECNN: cyberbullying detection in social media platforms using topic modelling and deep learning." *Multimedia Tools and Applications* (2023): 1-40.
[9] Neha, M. V., Sajan Muhammad, V. Indu, and Sabu M. Thampi. "Detection and Prevention of Cyberbullying in Social Media Using Cognitive Computational Analysis." In *Combatting Cyberbullying in Digital Media with Artificial Intelligence*, pp. 18-34. Chapman and Hall/CRC, 2023.
[10] Ahmed, Md Tofael, Almas Hossain Antar, Maqsudur Rahman, Abu Zafor Muhammad Touhidul Islam, Dipankar Das, and Md Golam Rashed. "Social Media Cyberbullying Detection on Political Violence from Bangla Texts Using Machine Learning Algorithm." *Journal of Intelligent Learning Systems and Applications* 15, no. 4 (2023): 108-122.
[11] Al-Ajlan, Monirah, and Mourad Ykhlef. "Firefly-cddl: A firefly-based algorithm for cyberbullying detection based on deep learning." *Computers, Materials & Continua* (2023).
[12] HUANG, HUANG, and DONGKAI QI. "Cyberbullying Detection on Social Media." *Higher Education and Oriental Studies* 3, no. 1 (2023).
[13] Fati, Suliman Mohamed, Amgad Muneer, Ayed Alwadain, and Abdullateef O. Balogun. "Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction." *Mathematics* 11, no. 16 (2023): 3567.