DETECTING CYBER BULLYING ON SOCIAL MEDIA IN THE AGE OF BIG DATA: DEEP LEARNING APPROACH

E. Babu[1], Bobbili Mounika[2], Sinthiya Sultana[2], Gosangari Rishitha[2], Nannaveni Pavan[2], Sainabarapu vm Siva Datta Prasad[2], Mandha Naresh[2]

[1]Assistant Professor, [2]UG Scholar, [1,2]Department of Computer Science and Engineering- (Cyber Security)
[1,2]Malla Reddy Engineering College and Management Sciences, Kistapur, Medchal,501401, Telangana.

Abstract

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 wilfully 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) 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. Many individuals, especially adolescents, suffer negative effects such as depression, sleeplessness, lowered self-esteem and even lack of motivation to live when being targeted by bullies on social media. Much is being done to stop regular bullying in schools. Traditional mechanisms to combat cyberbullying detections include the development of standards and guidelines that all users must adhere to, employment of human editors to manually check for bullying behavior, the use of profane synonym lists, and the use of regular expressions. However, these mechanisms fall short in social media. As a result, the maintenance of these mechanisms is time and labour consuming. Also, they cannot scale well. Therefore, it necessitates the use of a learning framework to accurately detect new cyberbullying detections automatically. Our research focuses on textual cyberbullying detection because text is the most common form of social media. However, the content information in social media is short, noisy, and unstructured with incorrect spellings and symbols, and this impacts the performance of some traditional machine learning methods based on vocabulary knowledge. For this reason, we propose a Char-CNN (Character-level Convolutional Neural Network) model to identify whether the text in social media contains cyberbullying. We use characters as the smallest unit of learning, enabling the model to overcome spelling errors and intentional obfuscation in real-world corpora.

Keywords: Social media, Cyberbullying, Artificial intelligence, Convolutional neural networks, N-gram feature selection.

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.

Our work proposes a Char-CNN (Character-level Convolutional Neural Network) model to identify whether the text in social media contains cyberbullying. This work proposes a new model with a character-level convolutional neural network to detect cyberbullying. Our model is essentially a classifier based on character-level convolutional neural network (CNN) with varying size filters. We use characters as the smallest unit of learning, enabling the model to learn character-level features to overcome the spelling errors and intentional obfuscation in data.

2. LITERATURE SURVEY

Traditional studies on cyberbullying stand more on a macroscopic view. These studies focused on the statistics of cyberbullying, explored the definitions, properties, and negative impacts of cyberbullying and attempted to establish a cyberbullying measure that would provide a framework for future empirical investigations of cyberbullying [7]. As cyberbullying has captured more attention, various methods have been used for the detection of cyberbullying in a given textual content. An outstanding work is the one by Nahar et al. Their work used the Latent Dirichlet Allocation (LDA) to extract semantic features, TF-IDF values and second-person pronouns as features for training an SVM [8].

Reynolds et al used the labelled data, in conjunction with the machine learning techniques provided by the Weka tool kit, to train a C4.5 decision tree learner and instance-based learner to recognize bullying content [9]. Xu et al showed that the SVM with a linear kernel using unigrams and bigrams as features can achieve a recall of 79% and a precision of 76% [6]. Dadvar et al took into account the various features in hurtful messages, including TF-IDF unigrams, the presence of swear words, frequent POS bigrams, and topic-specific unigrams and bigrams, and the approach was tested using JRip, J48, the SVM, and the naive Bayes [10].

Kontostathis et al analyzed cyberbullying corpora using the bag-of-words model to find the most commonly used terms by cyberbullies and used them to create queries [11]. In the work of Ying et al, the Lexical Semantic Feature (LSF) provided high accuracy for subtle offensive message detection, and it reduced the false positive rate. In addition, the LSF not only examines messages, but it also examines the person who posts the messages and his/her patterns of posting [12]. As the use of deep learning becomes more widespread, some deep learning-based approaches are also being used to detect cyberbullying.

The work of Agrawal and Awekar provided several useful insights and indicated that using learning-based models can capture more dispersed features on various platforms and topics [13]. The work of Bu and Cho provided a hybrid deep learning system that used a CNN and an LRCN to detect cyberbullying in SNS comments [14]. Since previous data-based work relied almost entirely on vocabulary knowledge, the challenge posed by unstructured data still exists. Some works observed that the content information in social media has many incorrect spellings, and in some cases, the users in social media intentionally obfuscate the words or phrases in the sentence to evade the manual and automatic detection [14]. These extra words will expand the vocabulary and affect the various performances of the algorithm.

Waseem and Hovy performed a grid search over all possible feature set combinations. They found that using character n-grams outperforms when using word n-grams by at least 5 F1-points using similar features [14], and it is a creative way to reduce the impacts of misspellings. Al-garadi et al used a spelling corrector to amend words, but we believe that some mistakes in this particular task scenario hide the speaker's intentions and correcting the spelling will destroy the features in the original dataset [15]. Zhang et al innovatively attempted to use phonemes to overcome deliberately ambiguous words in their work. However, some homophones with different meanings will get the same expression after their conversion, and their methods cannot solve some misspellings that have no association in their pronunciations [12].

Previous psychological and sociological studies suggested that emotional information can be used to better understand bullying behaviours, and thon emoticons in social text messages conveyed the emotions of users [13]. Dani et al presented a novel learning framework called Sentiment Informed Cyberbullying Detection (SICD), which leveraged sentiment information to detect cyberbullying behaviours in social media [14]. Unfortunately, in the past cyberbullying detection work, almost no work took into account these special symbols. As a common pre-processing technique, removing symbols and numbers destroys the features of the emojis in the original dataset.

We believe that spelling mistakes can be learned. Most of the spelling mistakes have an edit distance of less than 2, and there is a certain regular pattern, which is related to people's pronunciation habits and the key distribution on a keyboard [15]. In addition, on social networks, in order to convey a special meaning, some spelling mistakes are customary and common. Almost all factors suggest that these errors that we regarded as noise in previous works can be memorized by learning the combinations of characters. We use characters as the smallest unit since working on only characters has the advantage of being able to naturally learn unusual character combinations such as emoticons [15].

3. ROPOSED SYSTEM

The proposed architecture for cyberbullying detection as shown in Figure 3 is broadly divided into four stages namely data storage stage, data preprocessing stage, data detection stage and output stage. In the data storage stage, data will be trained based on word, character and synonyms. Finally creates the three individual trained datasets such as word level trained dataset, character level trained dataset and synonym level trained dataset. These trained data sources consisting of malicious data generated by numerous attackers and contains the spelling and grammatical errors, these datasets available from the different sources of social networking platforms.

3.1. Data preprocessing stage

In the data preprocessing stage, input test data (T) will be applied and will be spitted into words. Then white space will be removed using padding extraction operation. In the extracted words, there might be the special characters, unknown symbols, and encrypted format data. This may cause to creation of abusive content in text generates bullying. Thus, these missing unknown text data will be replaced by the known relevant text. The text data is in ASCII format generally, but neural networks neither be trained nor be tested with the text content. Thus, the input text data will be converted into special type of non-ASCII value and will be represented in digital numeric’s for every character like “**a** will be transformed to **0”**, similarly b:1, c:2, d:3 and goes on for all characters.

3.1.1 Tokenization

Over here the input text data is split into a set of words by removing all punctuation marks, tabs and other non-text characters and replacing them with white spaces. The part-of-speech (POS) tagging is also applied in some cases where words are tagged according to the grammatical context of the word in the sentence, hence dividing up the words into nouns, verbs, etc. This is important for the exact analysis of relations between words. Another approach was to ignore the order in which the words occurred and instead focus on their statistical distributions (the bag-of-words approach). In this case it is necessary to index the text into data vectors. The POS becomes important if the research is related to NLP. In one algorithm as part of extension work POS has been implemented.

3.1.2 Padding extraction

Padding refers to the white space between words, thus in padding extraction the space between two conjugative words will be extracted. In most of the times, the attackers wantedly use the whiter space to utilize the abusive text in the data. Thus, by using the padding extraction, the words contain white space will be precisely analyzed for cyberbullying detection.

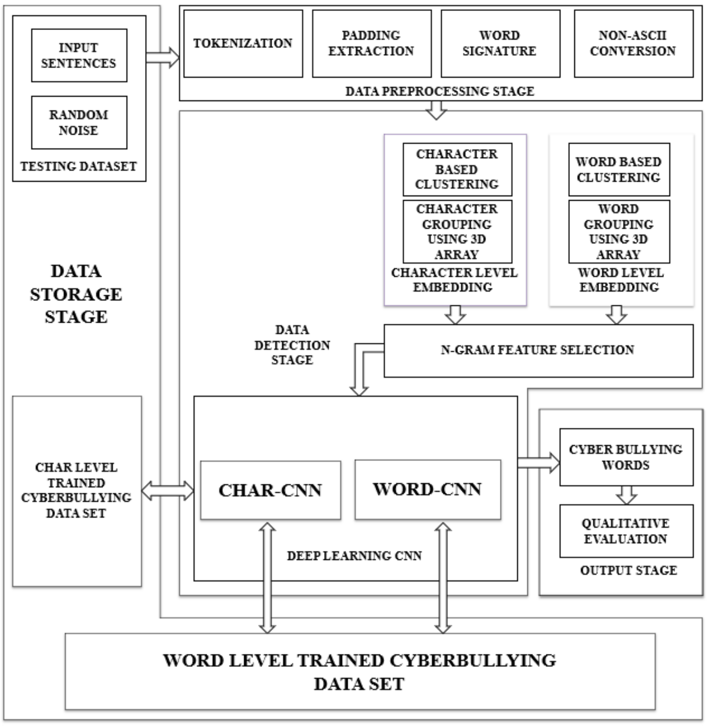


Figure 3: Proposed cyberbullying detection architecture.

3.1.3 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.

3.1.4 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.

4. RESULTS AND DISCUSSION

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.

MODULES

To implement this paper following modules are used:

1) Upload Dataset: Using this module text-based dataset can be uploaded to application.
2) Clean Module: Using this module we will apply various NLP techniques to remove stop words, special symbols etc.
3) Generate Vocabulary and Embedding Vector: using this module we will build vocabulary with all English characters set. Convert all text-based data to numeric by obtaining text numeric value from vocabulary and build a training vector.
4) Generate Character Based CNN Model: Using this module we will create CNN layers with vocabulary input and output sizes and then give train data as input to build CNN model.
5) Metrics Calculation: Using this module we will calculate various metric such as ACCURACY, PRECISION, RECALL and FMEASURE.
6) Predict Cyberbullying: Using this module we will ask user to enter any text message and then apply pre-processing technique to clean text and then convert text into one hot encoding or numeric vector. This numeric vector will be applied on CNN trained model to predict whether text contains any cyber bulling words or not.

Formulas to calculate metrics

Accuracy = correctly_classified_records / total_no_of_test_records

Precision = TruePositives / (TruePositives + FalsePositives)

Recall = TruePositives / (TruePositives + FalseNegatives)

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

While calculating accuracy suppose we have 20 test records and model able to predict 18 records correctly then the accuracy of model can be 18 / 20 = 0.90%. Similarly, to get Precision, Recall and F-Measure we need to calculate 4 values based on prediction.

TruePositives: this means model able to correctly classified/predicted given test record.

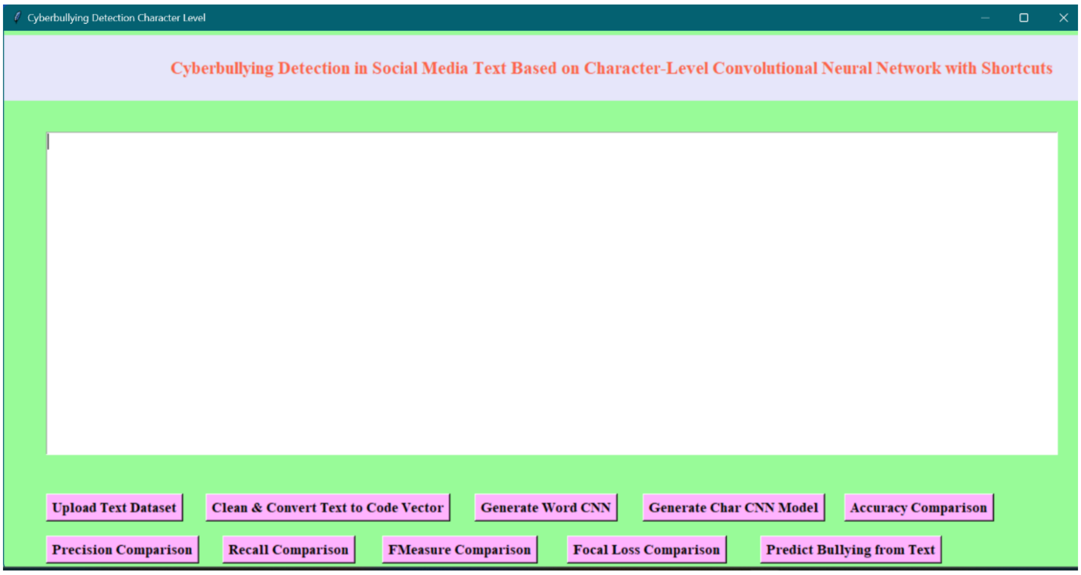
FalsePositives: this means model predicted given test record as positive, but it belongs to negative class.

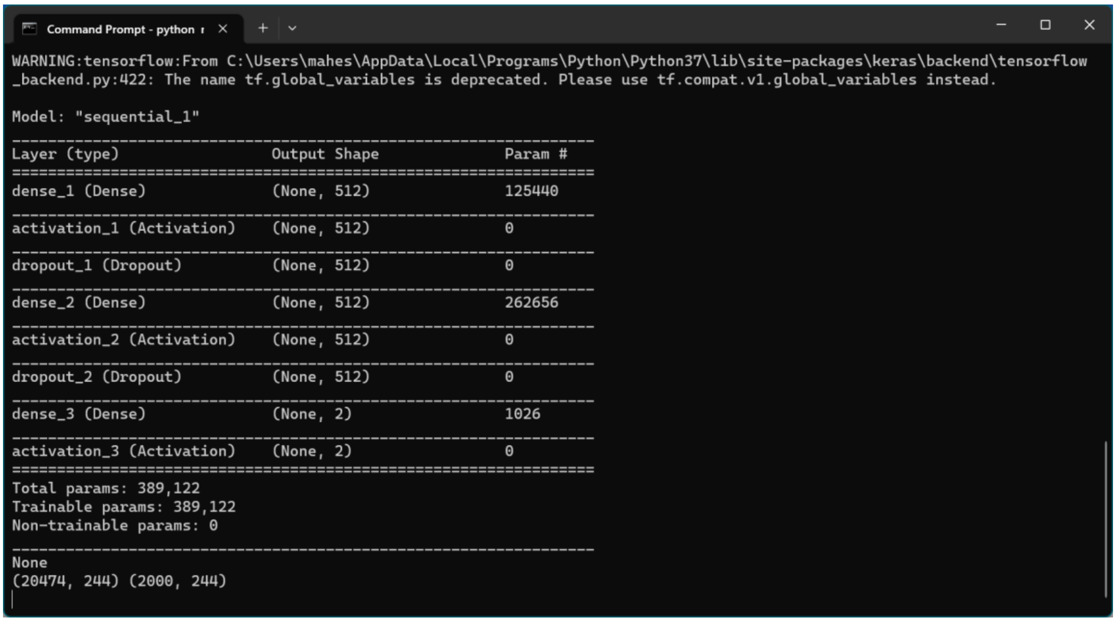
FalseNegatives: this means model predicted given test records as negative, but it belongs to positive class.

While prediction we will have count of TruePositives, FalsePositives and FalseNegatives based on these values we can calculate Precision, Recall and F-Measure.

Here we are building words and char based two CNN models and evaluating performance between them. To implement this project, we are using tweets dataset which contains more than 20000 records.

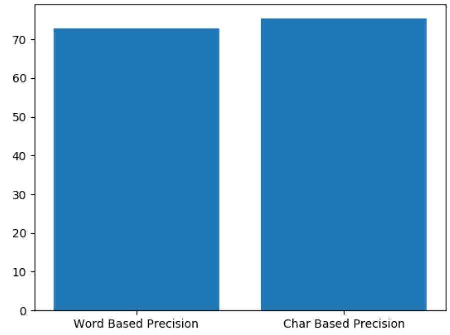
UI RESULTS

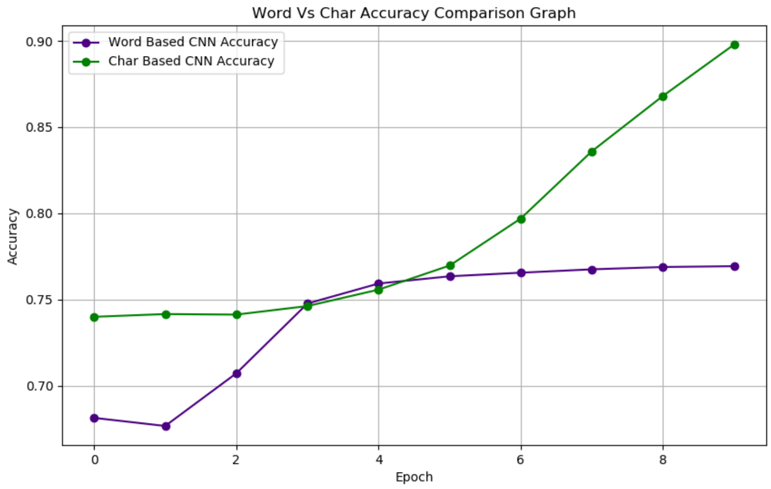

Cyberbullying Detection Character Level
Cyberbullying Detection in Social Media Text Based on Character-Level Convolutional Neural Network with Shortcuts
Upload Text Dataset
Clean & Convert Text to Code Vector
Generate Word CNN
Generate Char CNN Model
Accuracy Comparison
Precision Comparison
Recall Comparison
FMeasure Comparison
Focal Loss Comparison
Predict Bullying from Text

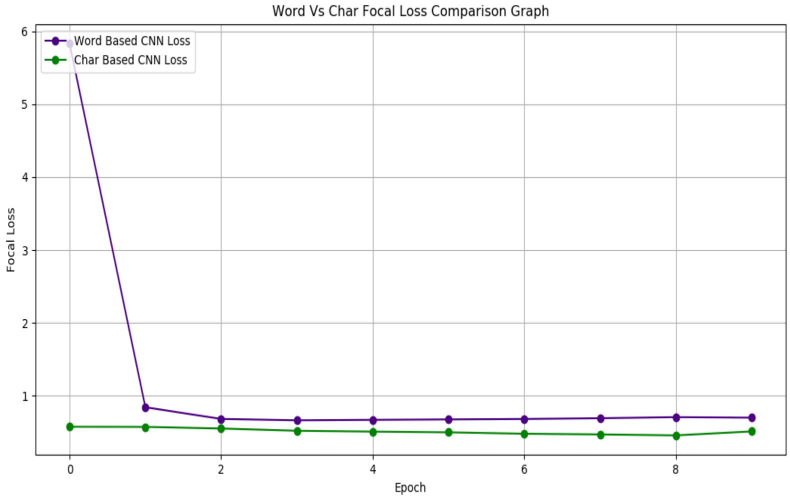


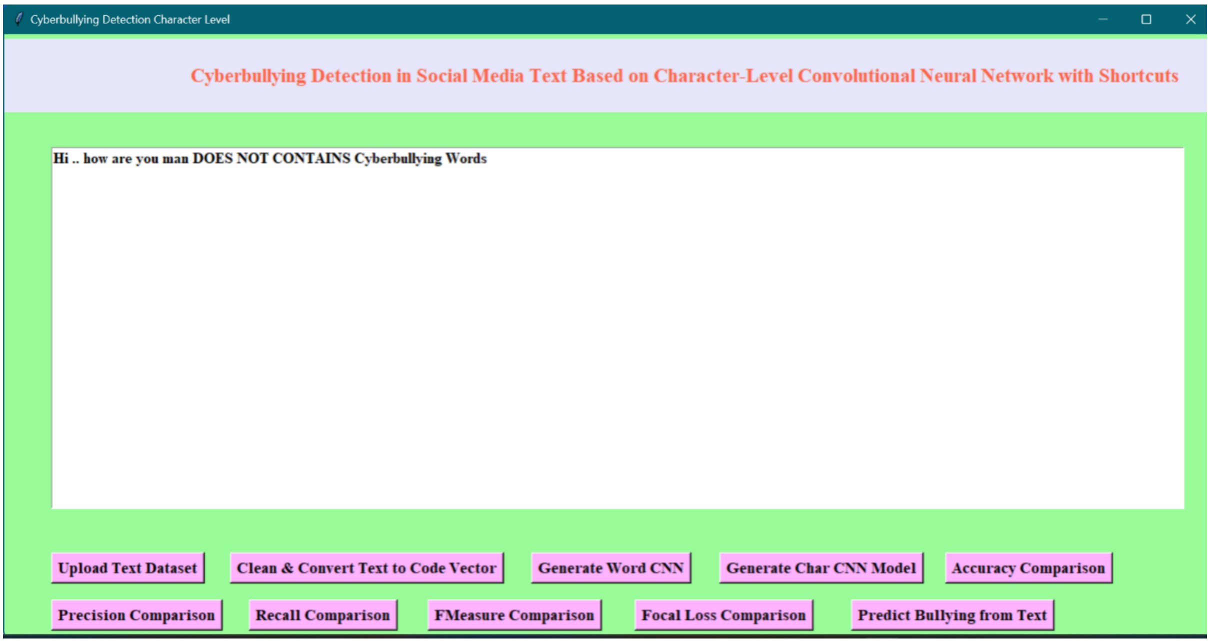
```
WARNING:tensorflow:From C:\Users\mahes\AppData\Local\Programs\Python\Python37\lib\site-packages\keras\backend\tensorflow
_backend.py:422: The name tf.global_variables is deprecated. Please use tf.compat.v1.global_variables instead.

Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 512)               125440
_________________________________________________________________
activation_1 (Activation)    (None, 512)               0
_________________________________________________________________
dropout_1 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 512)               262656
_________________________________________________________________
activation_2 (Activation)    (None, 512)               0
_________________________________________________________________
dropout_2 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_3 (Dense)              (None, 2)                 1026
_________________________________________________________________
activation_3 (Activation)    (None, 2)                 0
=================================================================
Total params: 389,122
Trainable params: 389,122
Non-trainable params: 0
_________________________________________________________________
None
(20474, 244) (2000, 244)
```


70
60
50
40
30
20
10
0
Word Based Precision
Char Based Precision

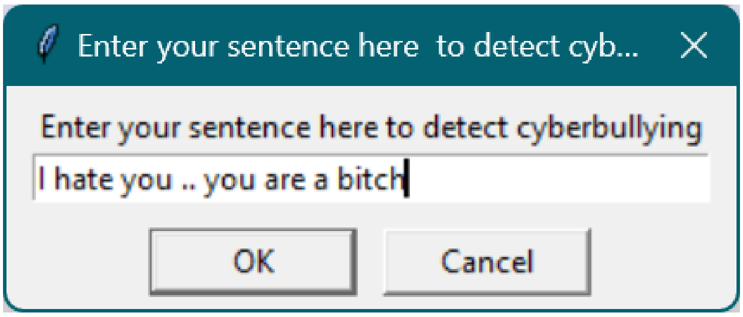

Word Vs Char Accuracy Comparison Graph
Word Based CNN Accuracy
Char Based CNN Accuracy
0.90
0.85
0.80
0.75
0.70
Accuracy
0
2
4
6
8
Epoch


Word Vs Char Focal Loss Comparison Graph
Word Based CNN Loss
Char Based CNN Loss
6
5
4
3
2
1
Focal Loss
0
2
4
6
8
Epoch


Cyberbullying Detection Character Level
Cyberbullying Detection in Social Media Text Based on Character-Level Convolutional Neural Network with Shortcuts
Upload Text Dataset
Clean & Convert Text to Code Vector
Generate Word CNN
Generate Char CNN Model
Accuracy Comparison
Precision Comparison
Recall Comparison
FMeasure Comparison
Focal Loss Comparison
Predict Bullying from Text

Hi .. how are you man DOES NOT CONTAINS Cyberbullying Words

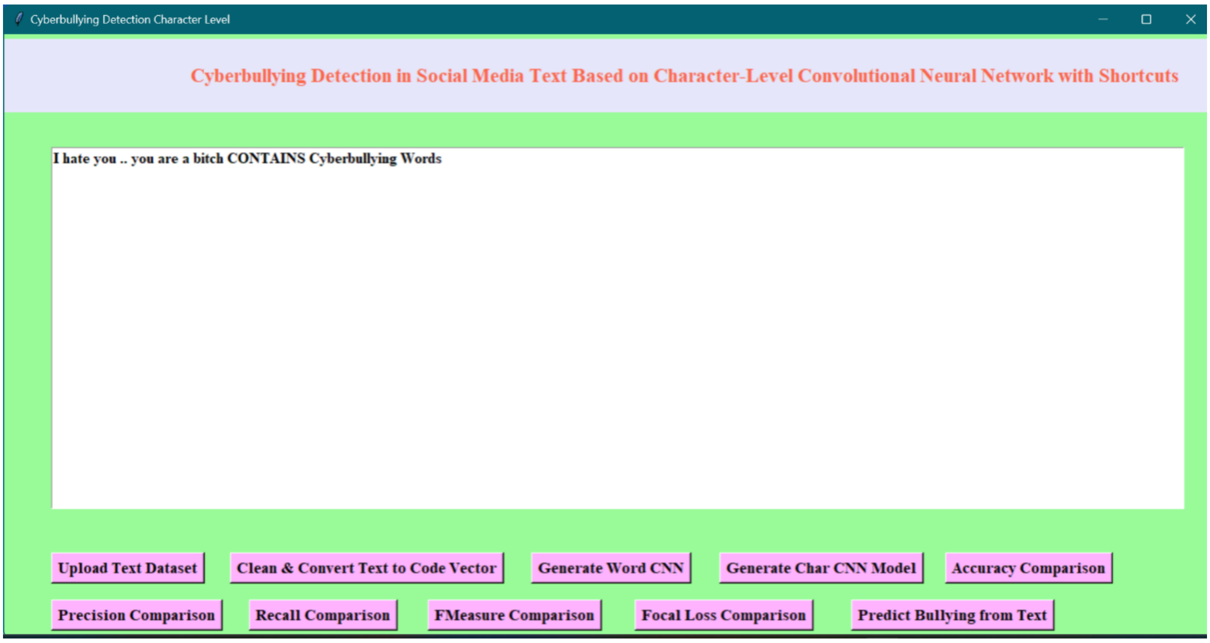
In above screen model predicted that given text message does not contain any cyber bullying words. Now will test with another sentence.



Enter your sentence here to detect cyberbullying

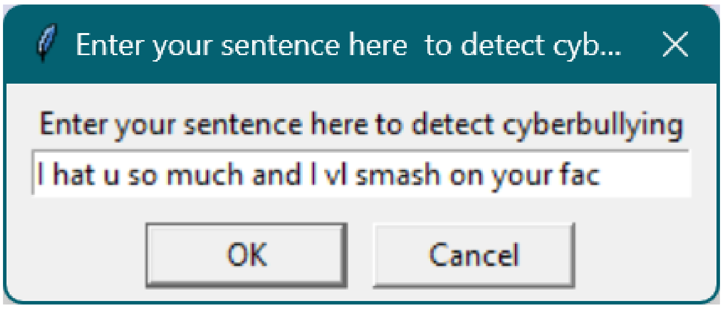
I hate you .. you are a bitch

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


Cyberbullying Detection Character Level
Cyberbullying Detection in Social Media Text Based on Character-Level Convolutional Neural Network with Shortcuts
Upload Text Dataset
Clean & Convert Text to Code Vector
Generate Word CNN
Generate Char CNN Model
Accuracy Comparison
Precision Comparison
Recall Comparison
FMeasure Comparison
Focal Loss Comparison
Predict Bullying from Text

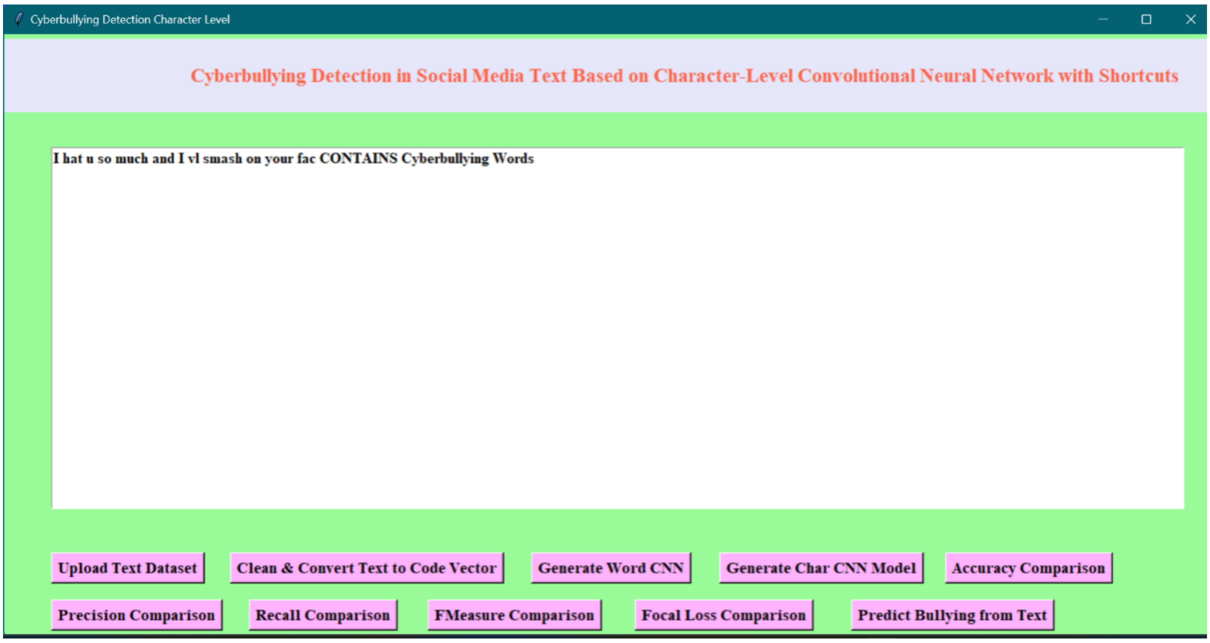
I hate you .. you are a bitch CONTAINS Cyberbullying Words

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



Enter your sentence here to detect cyberbullying

I hat u so much and I vl smash on your fac


Cyberbullying Detection Character Level
Cyberbullying Detection in Social Media Text Based on Character-Level Convolutional Neural Network with Shortcuts
Upload Text Dataset
Clean & Convert Text to Code Vector
Generate Word CNN
Generate Char CNN Model
Accuracy Comparison
Precision Comparison
Recall Comparison
FMeasure Comparison
Focal Loss Comparison
Predict Bullying from Text

I hat u so much and I vl smash on your fac CONTAINS Cyberbullying Words

5. CONCLUSIONS

In conclusion, cyberbullying is a serious issue that can have devastating effects on individuals and society as a whole. Therefore, effective methods for detecting and predicting cyberbullying are essential to prevent its harmful consequences. The use of deep learning models such as word-based and char based CNNs with shortcuts has shown promising results in accurately identifying instances of cyberbullying. These models utilize the power of natural language processing and deep learning techniques to analyze text data and classify it as either cyberbullying or non-cyberbullying. By incorporating shortcut connections, these models can learn complex features and relationships in the text data, enabling them to accurately predict and classify instances of cyberbullying.

Overall, the development of effective cyberbullying detection and prediction models is crucial in the fight against this harmful behavior. With further research and development, word-based and char based CNNs with shortcuts have the potential to become even more accurate and reliable in detecting and predicting cyberbullying, ultimately leading to a safer and more positive online environment for everyone.

REFERENCES

[1] StopBullying.gov. https://www.stopbullying.gov/

[2] Musu-Gillette L, Zhang A, Wang K, et al. Indicators of school crime and safety: 2017. National Center for Education Statistics and the Bureau of Justice Statistics. 2018.

[3] Hinduja S, Patchin JW. Bullying, cyberbullying, and suicide. Arch Suicide Res. 2010;14(3):206-221.

[4] Sugandhi R, Pande A, Chawla S, Agrawal A, Bhagat H. Methods for detection of cyberbullying: A survey. Paper presented at: 15th International Conference on Intelligent Systems Design and Applications; 2015; Marrakech, Morocco.

[5] Baldwin T, Cook P, Lui M, MacKinlay A, Wang L. How noisy social media text, how different social media sources. Paper presented at: 6th International Joint Conference on Natural Language Processing; 2013; Nagoya, Japan.

[6] Xu JM, Jun KS, Zhu X, Bellmore A. Learning from bullying traces in social media. Paper presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2012; Montreal, Canada.

[7] Freeman DM. Using naive Bayes to detect spammy names in social networks. Paper presented at: ACM Workshop on Artificial Intelligence and Security; 2013; Berlin, Germany.

[8] Reynolds K, Kontostathis A, Edwards L. Using machine learning to detect cyberbullying. Paper presented at: 10th International Conference on Machine learning and Applications and Workshops; 2011; Honolulu, HI.

[9] Kasture AS. A predictive model to detect online cyberbullying [master's thesis]. Auckland, New Zealand: Auckland University of Technology; 2015.

[10] Dadvar M, Ordelman R, de Jong F, Trieschnigg D. Improved cyberbullying detection using gender information. Paper presented at: 12th Dutchbelgian Information Retrieval Workshop; 2012; Ghent, Belgium.

[11] Dinakar K, Reichart R, Lieberman H. Modeling the detection of textual cyberbullying. Paper presented at: 5th International AAAI Conference on Weblogs and Social Media; 2011; Barcelona, Spain.

[12] Ying C, Zhou Y, Zhu S, Xu H. Detecting offensive language in social media to protect adolescent online safety. Paper presented at: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing; 2012; Amsterdam, Netherlands.

[13] Zhao R, Mao K. Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. IEEE Trans Affect Comput. 2017;8(3):328-339.

[14] Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2017;99:2999-3007.