Convolutional Neural Network based Cyberbullying in Social Media Detection Text based on Character level with shortcuts

Umarani Kunsoth, Sumitha Dhiravath

Department of Electronics and Communication Engineering

Sree Dattha Group of Institutions, Hyderabad, Telangana, India.

ABSTRACT

As people spend increasingly more time on social networks, cyberbullying has become a social problem that needs to be solved by machine learning methods. 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: convolutional neural networks, cyberbullying detection, social network, text classification

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 [15-18].

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 [19].

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 [8]. 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 [20].

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 [21]. 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 [22].

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 [23, 24]. 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 [25], 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 [26]. 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 [24]. 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 [27].

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 [23]. 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 [28, 29]. 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 [30].

3. PROPOSED 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.

3.1.1Tokenization

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.

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.

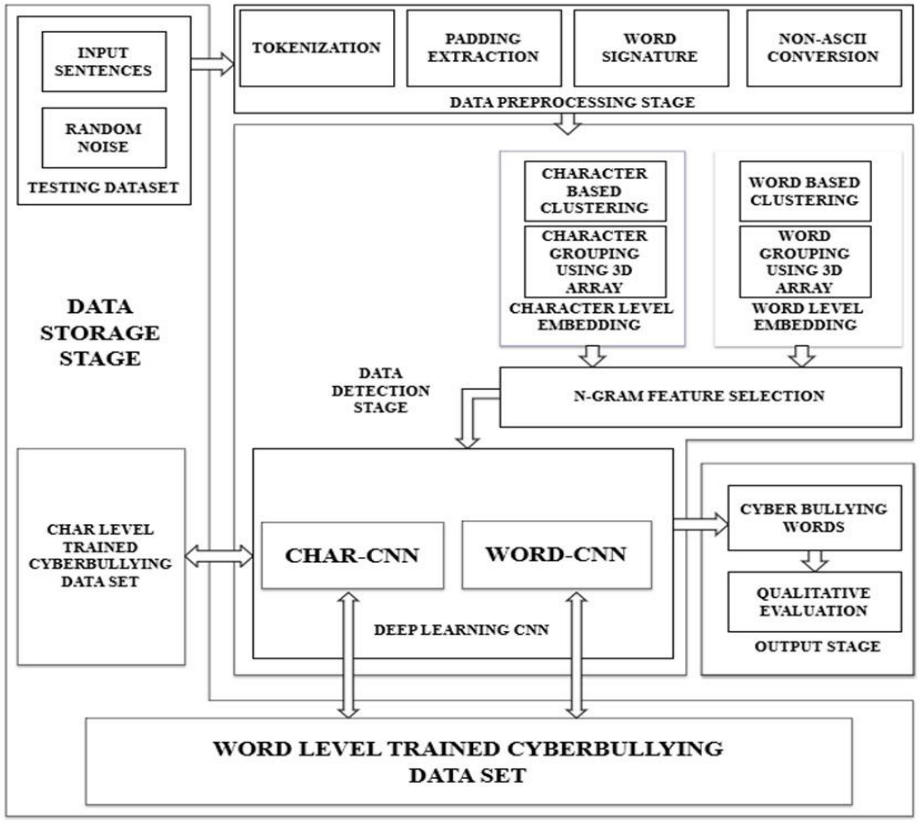


Fig. 1: Proposed cyberbullying detection architecture.

3.2 Data detection stage

In the data detection stage character level, word level and synonym level embedding operation will be performed. In this embedding character recognition, word recognition and synonym recognition operations will be performed parallel manner to give the maximum efficiency to detect the cyberbullying. Then the data groups will be formed as 3D array using pattern matching operations. The selection of character level or word level or synonym level cyberbullying detection is performed by the user through user interface. Then corresponding 3d group array will be applied CNN.

3.2.1 Data Clustering

Clustering is a powerful and broadly acceptable data mining technique which is used to partition voluminous data into different classes, known as clusters, to support the businessman or an end user by providing different views and various patterns of same data suitable to the requirements. The cyberbullying detection focuses on the different levels clustering's such as character level, word level and synonym-character level.

Phase I: The set of prototype vectors are much higher than the expected number of clusters. The prototypes are grouped to form the actual character-based clusters.

Phase II: In this phase word level clustering algorithm is executed on the prototypes vectors to find clusters' word centroid. Clustering of vast number of words in text samples is a key process in providing a higher level of knowledge about the underlying inherent classification of the abusive content causes to create cyberbullying.

Phase III: The word-based cluster centres obtained in Phase II are used in phase III. Synonym-character identification algorithm utilizing the high standard vocabulary is applied in this phase to generate the actual synonym-character-based clusters. The result from word level clustering which is found in phase II is used as the initial seed of the Synonym-character identification algorithm. Phase III converges quickly when the centroids from Phase II are used.

3.2.2 Grouping using 3D array

A normalized longest common subsequence (NLCS) based string approximation method is proposed for indexing multidimensional data cube. In this indexing system, the reference table is made, and dimension al key values are stored for each dimension. A dimensional reference table is a set of dimensional key values stored in sorted order. The slot number of a key value in the dimensional reference table will be the index of the key value on the axis of multidimensional array. NLCS based string approximation is used to search a nearest keyword for a misspelled keyword, in the reference table and gets its slot number.

Normalized LCS based string approximation is used to design a character, word, and synonym (CWS) searching algorithm. This CWS searching algorithm gives near optimal solution to the string approximation problem. The algorithm finds the NLCS values of searched keyword with all the stored keywords in the set. The keywords in the set having NLCS value between 0.5 and 1 are the nearest neighbor of the searching keyword. The keyword closest to the searching keyword having highest NLCS value will be the optimal keyword. The CWS searching, finds the index of keyword, like searching keyword from the set of stored keywords and creates the 3d array group for easily detection of cyberbullying. So, the abusive words and its synonyms will be identified easily.

3.2.3 N-gram Feature selection

The N-gram model combined with latent representation on the data classification task. Their model called as supervised n-gram embedding uses a multi-layer perceptron to accomplish the embedding. The number of distinct character and word-based N grams in a text can be as high and its feature selection vector size extremely high even for moderate values of n. The N-gram Feature selection applied only on character and word-based embedding vectors as it does not apply on synonym based embedding vector. Because synonym-based vectors are classified initially in the synonym level embedding so there is no requirement to generate the features again. If the N-gram feature selection

applied on synonym based embedding vectors, then classification accuracy will reduce because of original features will get loosed. However, only a small fraction of all possible character and word-based n grams will be present in any given set of documents, thereby reducing the dimensionality substantially. The dimensionality reduction problem is handled in the present work in two different approaches where one set of n grams are identified as valid N grams and other set is treated as invalid N grams. The adequacy of this model is also evaluated in terms of average information conveyed by valid N grams in comparison with invalid N grams.

3.3 Deep-learning CNN

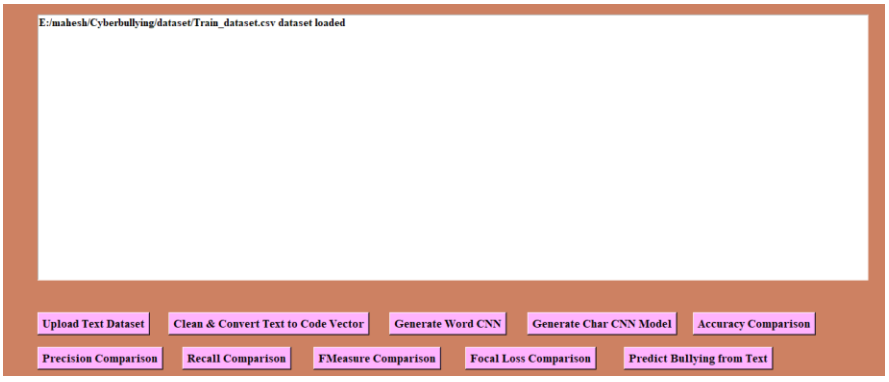
The selection of character level (C) or word level (W), synonym level (S) cyberbullying detection is performed by the user through user interface, and then selected cyberbullying detection operation will be performed on the predefined trained model. There is common deep learning architecture for word, character and synonym-character level cyberbullying detection, if the user selects word level cyberbullying detection, then entire operation will be performed on word level trained dataset. If the user selects character level cyberbullying detection, then entire operation will be performed on character level trained dataset. But, if the user selects synonym-character level cyberbullying detection, then entire operation will be performed on three trained datasets such as character level, word level and synonym-character level trained datasets. The CNN can take predefined features of synonyms and n-grams features of each word and character, the applied to multiple layers of various filters to build feature vector. The advantage of this technique is each word, character and synonym similarity will be checked with CNN trained model and if attacker is using more cyberbullying content then it can be easily detected. With the help of this technique any spelling variations used by attacker to avoid detection also detected very easily. Apart from spelling mistake malicious users, we are building CNN to one more extra step by utilizing similar synonym words of cyberbullying text to mislead and missed detector words as detector trained model with original words and also with various similar synonyms of original word. Users can understand that text contains cyberbullying with synonyms words, but model do not know. Thus, we are building CNN model with synonyms of all possible ways for detection with characters to prevent malicious user from sending cyberbullying text in no possible way. Finally, in the output stage the detected cyberbullying words (CB) will be generated. If the CNN model generates CB as 1 it indicates cyberbullying word else generates CB as 0 it indicates non-cyberbullying word. Then qualitative evaluation operation will be performed to measure the efficiency of the system. The detailed algorithm for proposed cyberbullying detection is presented in Table 1 using CNN model.

Table 2: CNN based cyberbullying detection algorithm.

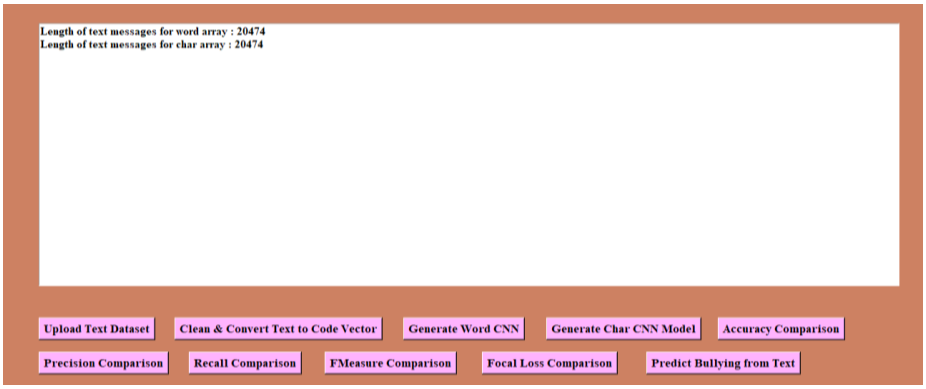
Inputs: T, F, K, P, W, C, S Output: CB
Begin
Step 1: Tokenization of input sentences
Step 2: Remove white space through Padding Extraction
Step 3: Calculate the total number of Tweet words T

Step 4: Word to Non Ascii conversion
Step 5: Word, character and synonym-character level clustering→ { WC, CC, SC }
for $\boldsymbol{i = 1:T}$ begin
Step 6: 3D-array grouping of WC, CC and SC → {$3D-W, 3D-C, 3D-S$ }
Step 7: N-gram feature selection on {$3D-W, 3D-C$ } → {$3D-NG-WC$}
Step 8: Apply CNN using {$3D-NG-WC, 3D-S$ } {F_W, F_C, F_S} =Fullyconnected2D ($3D-NG-WC, 3D-S, F, K$) M_W=MaxPooling2D (W, F_W, P, K) C_W=Conv2D(W, M_W, F, K) M_C= MaxPooling2D (C, F_C, P, K) C_C= Conv2DW layer(C, M_C, F, K) M_S= MaxPooling2D (S, F_S, P, K) C_S= Conv2D (S, M_S, F, K) CB=Embedding2D(C_W, C_C, C_S, F, K) end end

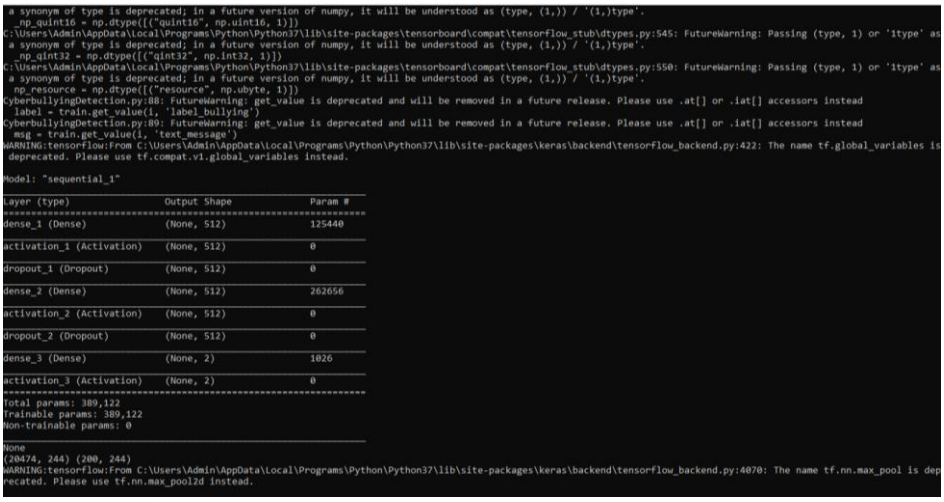
4. RESULTS



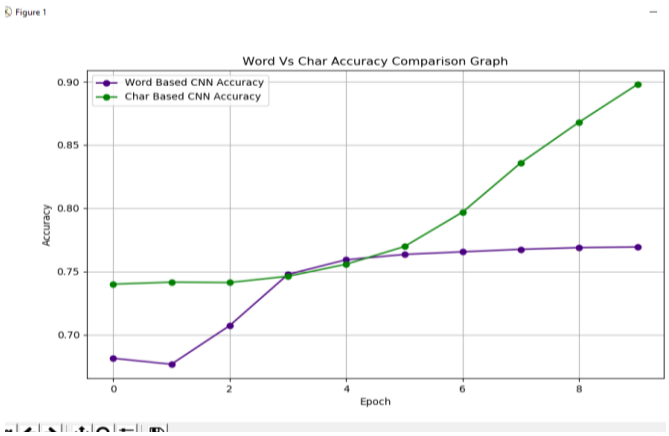
Now click on 'Clean & Convert Text to Code Vector' button to read text and convert to vector using vocabulary index



In above screen after cleaning and converting text to vector we got 20474 tweets records and now click on 'Generate Word CNN' button to embed all tweets to vocabulary index and then generate train data and build word CNN model.

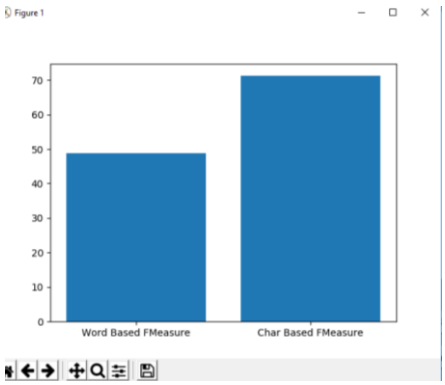


In above screen CNN different layers created with different size to filter data and to train model. Now click on 'Generate Char CNN Model' button to generate model using characters

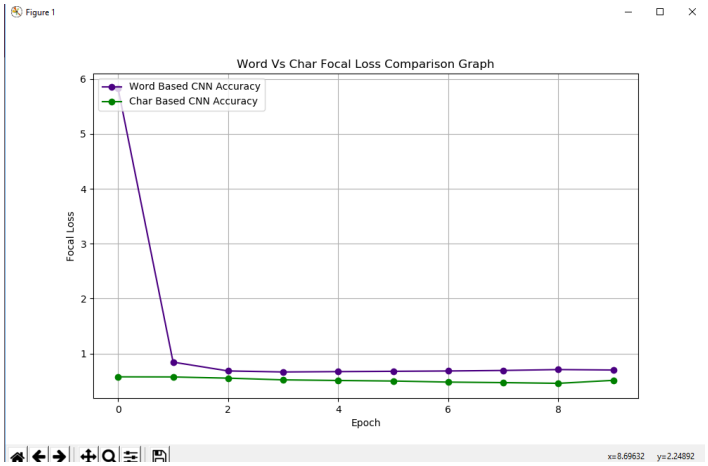


In above screen to generate model we used EPOCH as 10 iterations and in above graph x-axis represents epoch value and y-axis represents accuracy at each epoch iteration. In above graph blue line indicates word based and green line indicates CHAR based CNN accuracy.

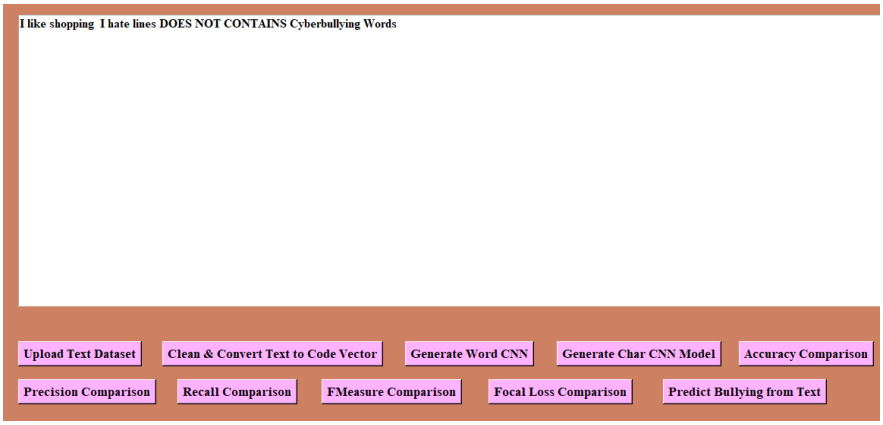
Below is F-Measure graph



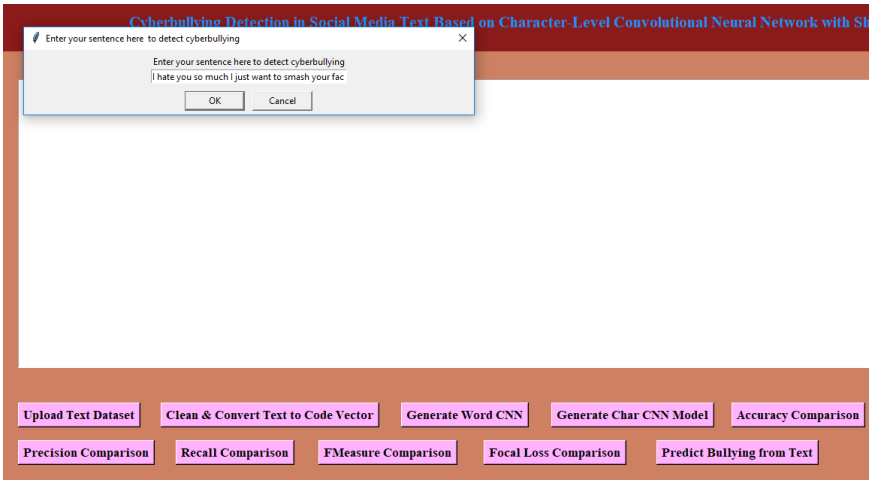
Below is loss graph



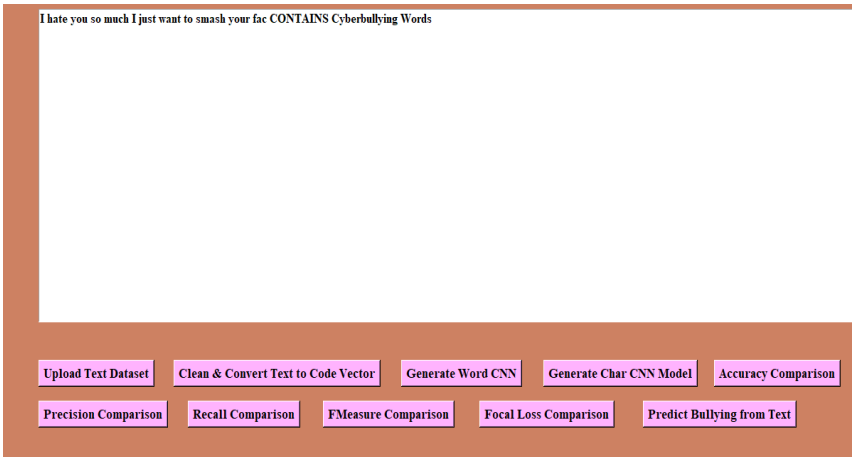
In above loss graph char-based CNN has less loss compare to word based CNN. Now click on 'Predict Bullying from text' button to allow user to enter some text and then Char based model will give prediction result



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



In above screen I entered message as 'I hate you so much I just want to smash your fac' and in that sentence I done spelling mistake to face as fac and below is the result



In above screen we got prediction result as given message contains Cyber Bullying words.

4. CONCLUSION AND FUTURE WORK

We provide a new Chinese Weibo comments dataset of 19K comments specifically for cyberbullying detection. All the samples belonging to more than 20 celebrities who have bad reputations or who have experienced some vicious incidents have been selected and manually annotated. These data are ideal for experiments on identifying bullying content, because these public figures have the potential to become cyberbullying victims. We also propose an automatic solution to identify whether the text in social media contains cyberbullying. It learns the char-level features to overcome spelling errors and intentional obfuscation in data. Shortcuts are utilized to stitch different levels of features to learn hybrid bullying signals, and we adopt a focal loss function to overcome the class imbalance problem in the dataset. These well-designed modules are simple but truly effective, and our approach has better performance with a P, F, and R of 79.0, 71.6, and 69.8, respectively, on the Weibo dataset and 81.0, 74.2, and 70.5, respectively, on the Tweet dataset. Compared with other schemes, the Char-CNNS with the focal loss performs relatively well for different data distributions. The experimental results demonstrates that the character-level embedding and hybrid features from multiple layers increase the performance of cyberbullying detection on social media text. The experiment shows that a shallow neural network model already works excellently, and it has achieved satisfactory results. In the future,

we hope to expand its number of layers to explore if there is any further improvement. It is also a worthwhile attempt to adjust the weights of the shortcut branches. In addition, identifying more types of bullying is also a direction that we want to explore in the future.

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.
[15] Patchin JW, Hinduja S. Bullies move beyond the schoolyard a preliminary look at cyberbullying. Youth Violence Juvenile Justice. 2006;4(2):148-169.
[16] Robert S, Smith PK. Cyberbullying: another main type of bullying? Scand J Psychol. 2008;49(2):147-154.
[17] Smith PK, Jess M, Manuel C, Sonja F, Shanette R, Neil T. Cyberbullying: its nature and impact in secondary school pupils. J Child Psychol Psychiatry. 2008;49(4):376-385.

[18] Tokunaga RS. Following you home from school: a critical review and synthesis of research on cyberbullying victimization. Comput Hum Behav. 2010;26(3):277-287.
[19] Nahar V, Xue L, Pang C. An effective approach for cyberbullying detection. Commun Inf Sci Manag Eng. 2013;3(5):238-247.
[20] Kontostathis A, Reynolds K, Garron A, Edwards L. Detecting cyberbullying: Query terms and techniques. Paper presented at: 5th Annual ACM Web Science Conference; 2013; Paris, France.
[21] Agrawal S, Awekar A. Deep learning for detecting cyberbullying across multiple social media platforms. Paper presented at: 40th European Conference on IR Research; 2018; Grenoble, France.
[22] Bu SJ, Cho S. A hybrid deep learning system of CNN and LRCN to detect cyberbullying from SNS comments. Paper presented at: International Conference on Hybrid Artificial Intelligence Systems; 2018; Oviedo, Spain.
[23] Dani H, Li J, Liu H. Sentiment informed cyberbullying detection in social media. Paper presented at: Joint European Conference on Machine Learning and Knowledge Discovery in Databases; 2017; Skopje, Macedonia.
[24] Zhang X, Tong J, Vishwamitra N, et al. Cyberbullying detection with a pronunciation based convolutional neural network. Paper presented at: 15th IEEE International Conference on Machine Learning and Applications; 2016; Anaheim, CA.
[25] Waseem Z, Hovy D. Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. Paper presented at: North American Chapter of the ACL Student Research Workshop; 2016; San Diego, CA.
[26] Al-garadi MA, Varathan D, Ravana SD. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Comput Hum Behav. 2016;63:433-443.
[27] Kokkinos CM. The relationship between bullying, victimization, trait emotional intelligence, self-efficacy and empathy among preadolescents. Soc Psychol Educ. 2012;15(1):41-58.
[28] Damerau FJ. A technique for computer detection and correction of spelling errors. Commun ACM. 1964;7(3):171-176.
[29] Kemighan MD, Church KW, Gale WA. A spelling correction program based on a noisy channel model. Paper presented at: 13th International Conference on Computational Linguistics; 1990; Helsinki, Finland.
[30] Zhang X, Zhao J, Lecun Y. Character-level convolutional networks for text classification. Paper presented at: International Conference on Neural Information Processing Systems; 2015; Montreal, Canada.
[31] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. Paper presented at: 25th International Conference on Machine Learning; 2008; Helsinki, Finland.
[32] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. J Mach LearnRes. 2011;12(1):2493-2537.
[33] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278-2324.
[34] Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882. 2014.
[35] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188. 2014.

[36] Severyn A, Moschitti A. Twitter sentiment analysis with deep convolutional neural networks. Paper presented at: 38th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2015; Santiago, Chile.
[37] Johnson R, Tong Z. Convolutional neural networks for text categorization: Shallow word-level vs. deep character-level. arXiv preprint arXiv:1609.00718. 2016.
[38] Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. Comput Sci. 2012;3(4):212-223.
[39] Raisi E, Huang B. Weakly supervised cyberbullying detection with participant-vocabulary consistency. Soc Netw Anal Min; 8(1).
[40] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Paper presented at: 13th International Conference on Artificial Intelligence and Statistics; 2010; Sardinia, Italy.