Unveiling Cyberbullying with a Character-CNN Model and Data Grouping for Enhanced Detection

Kondragunta Rama Krishnaiah, Professor, Department of Computer Science and Engineering, R K College of Engineering, Vijayawada - 521456, Andhra Pradesh, India, email: kondraguntark@gmail.com

Alahari Hanumant Prasad, Professor, Department of Computer Science and Engineering, R K College of Engineering, Vijayawada - 521456, Andhra Pradesh, India, email: hanuma.alahari@gmail.com

ABSTRACT

In today's digital age, the rise of social media has brought with it a troubling issue - cyberbullying. It has become a pressing social problem that calls for effective solutions, and one promising approach lies in the realm of machine learning. Our research focuses on tackling textual cyberbullying, as it is the most prevalent form of communication on social networks. However, dealing with social media content presents unique challenges. The texts are often short, noisy, and lack a structured format. Additionally, they may contain misspellings, symbols, and intentional distortions, making it difficult for traditional machine learning methods that heavily rely on vocabulary knowledge to perform well. To address these obstacles, we introduce a novel approach known as the Char-CNN (Character-level Convolutional Neural Network) model. Unlike traditional methods that work with words, our model operates at the character level. This means that the model learns from individual characters rather than entire words, making it more robust in handling spelling errors and intentional obfuscation present in real-world social media data. By leveraging the power of Char-CNN, we aim to accurately identify instances of cyberbullying in social media texts. This advancement holds significant promise in curbing the harmful effects of cyberbullying and creating a safer and more positive online environment for all users.

Keywords: Cyberbullying, social media, convolutional neural networks, word-CNN, char-CNN.

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. RELATED WORK

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. SYSTEM IMPLEMENTATION

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'

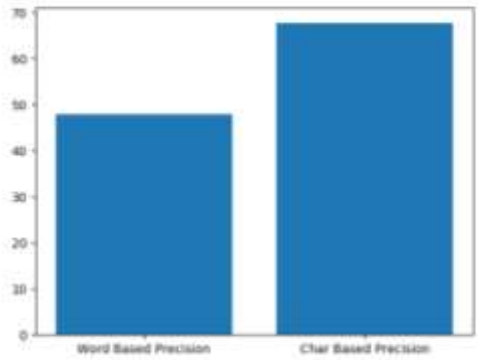
If user give input as 'bc' then CNN convert 'b, c' with embedding weight such as '1,3' as b is available at index 1 and c available at index 2. Similarly, CNN will build model by scanning embedding vocabulary.

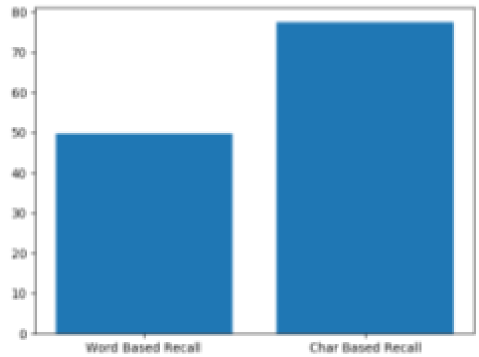
4. IMPLEMENTATION AND RESULTS

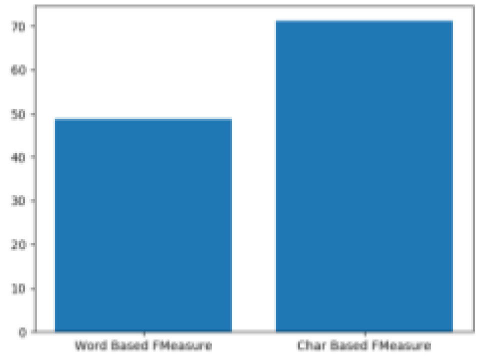
To implement this paper following modules are used:

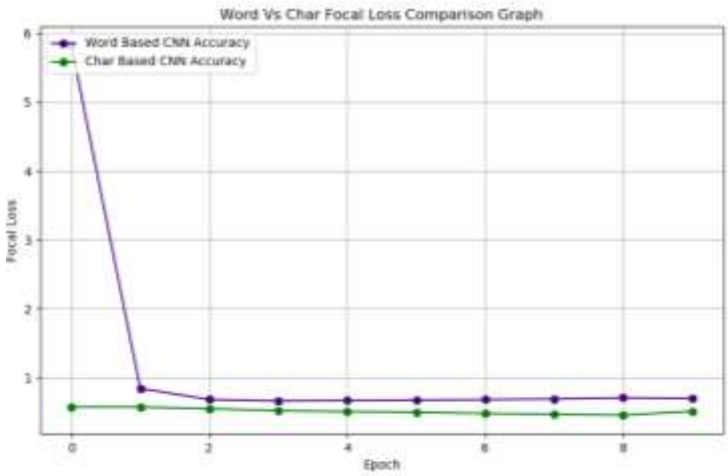
1) Upload Dataset: Using this module text-based dataset can be uploaded to application.
2) Clean Module: Using this module we will apply various NLP techniques to remove stop words, special symbols etc.
3) Generate Vocabulary and Embedding Vector: using this module we will build vocabulary with all English 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.

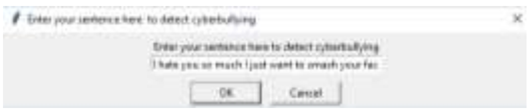
Here we are building words and char based two CNN models and evaluating performance between them. To implement this project, we are using tweets dataset which contains more than 20000 records.











5. CONCLUSION

This paper focused 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, a Char-CNN model is proposed to identify whether the text in social media contains cyberbullying. In addition, characters are used as the smallest unit of learning, enabling the model to overcome spelling errors and intentional obfuscation in real-world corpora.

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.