Cyber Bullying Detection on Twitter Mining

Saloni Wade[1], Maithili Parulekar[2], Prof. Kumud Wasnik[3]

[1,2,3]*Computer Science and Technology UMIT, SNDT Women's University* Mumbai, India

---

***Abstract*—**Social Media which is used as a platform of en-tertainment, job opportunity, marketing has also led to cyber-bullying. Social networking sites provide a prolific medium for autocrats and teens who use these sites and are susceptible to attacks. Cyberbullying is when technology is used as a medium to bully someone. Cyberbullying includes insulting, humiliating and making fun of people on social media that can cause mental breakdowns, it can affect one physically as well to the extent that can also lead to suicidal attempts. Through deep learning, we can detect patterns used by bullies and develop principles to automatically recognize cyberbullying content. Existing works for cyberbullying identification have at least one of the following bottlenecks. First, they address just one topic of cyberbullying. Second, they rely only on the particular characteristics of the data. We intend to show that deep learning-based models can overcome all the bottlenecks. Knowledge learned by these models on one dataset can be transferred on to other datasets. We performed thorough tests using real-world datasets: Twitter. We also provide a method via which we can get to know the percentage of positive, negative and neutral tweets.

***Index Terms*—Cyberbullying, Deep Neural Network(DNN) , Convolutional Neural Network(CNN),Twitter**

1. INTRODUCTION

With the progression of technology, the fad of social net- working platforms is increasing. Online users now share their information easily using computers and mobile. However, this has led to the growth of cyber criminal acts like cyberbul-lying which has become a global epidemic. Cyberbullying is the use of electronic communication to bully a person by sending objectionable messages using social media, immediate messaging or through digital messages. The main problem in obstructing cyberbullying is detecting its existence so that suitable action can be taken at the beginning stages. To overcome this problem, many methods and techniques had been worked upon until now to control this problem. Studies show that about 18 percent of the children in Europe have been involved in cyberbullying. Cyberbullying needs to be known and approached from different perspectives. Automatic disclosure and prevention of these incidents can substantially help to stop this problem. Moreover, most of the online platforms which are commonly used by teenagers have safe centers, for example, YouTube Safety Centre4 and Twitter Safety and Security 5, which provide support to users and monitor the communications. Most of the existing studies have used conventional Machine Learning (ML) models to detect cyberbullying incidents. Recently Deep Neural Network Based (DNN) models have also been applied for the detection of cyberbullying. Based on their reported results, their models outperform traditional ML models, and most important authors have stated that they have applied transfer learning which means their developed models for discovery of cyberbullying can be adapted and used on other datasets. Cyberbullying takes place in almost all of the online social networks; therefore, developing a detection model that is adaptable and transferable to different social networks is of great value. We extend our work by re-implementing the models on a new dataset. For this purpose, we have used a Twitter dataset that has been extensively used in cyberbullying studies. This provides a base to compare the outcome of DNN models with the conventional ML models.

2. RELATED WORK

Cyberbullying is recognized as a phenomenon at least since 2003. The use of social media exploded with the launching of multiple platforms such as Wikipedia (2001), MySpace (2003), Orkut (2004), Facebook (2004), and Twitter (2005). By 2006, researchers had pointed out that cyberbullying was a serious phenomenon as offline bullying. However, the au- tomatic discovery of cyberbullying was addressed only since 2009. As an examination topic, cyberbullying discovery is a text classification problem. Most of the existing works fit in the following template: get training dataset from single SMP, engineer a variety of characteristics with a certain style of cyberbullying as the objective, apply a few traditional machine learning methods, and evaluate success in terms of measures such as F1 score and accuracy. These works heavily rely on handcrafted features such as the use of swear words. These methods tend to have low accuracy for cyberbullying detection as handcrafted features are not robust against variations in bullying style across SMPs and bullying topics. Only recently, deep learning has been applied for cyberbullying discovery.

3. LITERATURE SURVEY

1. Andrew M. Dal and Quoc V. Le proposed a supervised sequence learning model using CNN and LSTM. The semi-supervised learning is the combination of supervised and unsupervised learning where by using this the unlabeled data is proved to be more useful in improving the generalization of the subsequent supervised model. The paper recommends the use of LSTM-RNN to be more useful than CNN and RNN for the purpose of data training using the proposed approach. This paper tests the semi-supervised method on five benchmarks to check the results using LSTM as the training method. This paper proves that CNN-LSTM is the better method than conventional CNN and gives better results than the previous methods for training unlabeled data.

2. H. Zeng, et. al. used a visualization technique that has 4 linked view that's helps to analyze learning parameters. These study uses AlexNet as the neural network architecture. It is necessary to get the insight of how the model parameters evolve from lower to higher accuracy so that we can improve the training process. The two main challenges in exploring the relationship between model parameters and performance

i.e Scalability and Interpretability are solved here. Various parameters and activation values between two CNN snapshots are evaluated based on TFlearn framework. As training process of CNN leads to large number of parameters over time, this results in decreased performance. This paper helps to view the learning process of the CNN.

3. Rui Zhao and Kezhi Mao used a new representation learning method have been proposed to tackle this problem. This method is Semantic-Enhanced Marginalized Denoising AutoEncoder (smSDA) developed via semantic extension of the popular deep learning model stacked denoising autoen- coder. The semantic extension consists of semantic dropout noise and sparsity constraints, where the semantic dropout noise is designed based on domain knowledge and the word embedding technique.

4. K. Sahay, et. al. explains that online bullying and aggression against social media users have grown abruptly. The research experimenting with different work process makes a robust methodology for extracting text, user experimenting with different methods the work process a robust methodology for extracting text, user work in certain ways to identify and classify bullying in the text by analyzing and network-based attributes studying the properties of bullies and aggressor and what feature distinguish them for regular user. The NLP and machine learning are studied and evaluated for the task of identifying bullying comments in the dataset. This paper shows the training in machine learning model using supervised learning

5. V. N. Kumar, et. al. proposed that the effective represen- tation of content is necessary for proper learning. This paper use naïve Bayes as the classifier for the content classification in email application it deals with the classification of spam words when massage is received and it is processed using feature set extraction method in which feature probabilities are found using NB and SVM are compared for precision factor .this paper just classifies the massage into cyberbullying. the denoised value for each word is calculated by grouping massage .this system alerts the system. It uses word embedding. technique which obtains bullying character automatically .the various modules use in this paper are GUI designing ,training dataset, classification and analyzing the twitter massages.

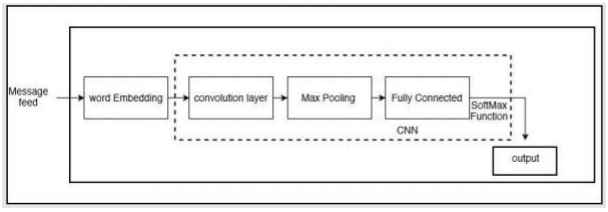

Fig. 1. System Flow

4. PROPOSED SYSTEM

A. *Message feed*

The input for the system consists of message feed from dataset. This is the input for word embedding and beginning of the workflow in the system.

B. *Word Embedding*

The data from the message feed is embedded into numerical form for the input of the CNN. Each word is represented by a real value vector. The distributed representation of words is learned by the technique of transfer learning.

C. *Convolution Neural Network*

Vectors generated by word embedding is the input for the neural network layers. Convolution layer is the first layer of CNN output of this layer is given to the max-pooling layer and fully connected network is generated. Softmax function is used after the fully connected layer to generate the output.

5. DEEP NEURAL NETWORK MODELS USED (DNN)

In this study, two different DNN models were used for the detection of cyberbullying: Convolutional Neural Network (CNN)and Long Short-Term Memory (LSTM).

These models respectively differ in complexity in their neural architecture. CNN's are usually used for image and text classification as well as sentiment classification. LSTM networks are used for learning long-term dependencies. Their internal memory makes these networks beneficial for text classification. All the models have similar layers except for the neural architecture layer which is unique to each model. The embedding layer, which will be explained in more detail in the following section, processes a fixed-length sequence of words. Then there is the fully connected layer which is a dense output layer with the abundance of neurons equal to the number of classes. The outermost layer is the softmax layer that provides softmax activation. Every model is trained using backpropagation with Adam optimizer and categorical cross-entropy loss function.

A. Convolutional Neural Network

Convolutional Neural Networks Convolutional Neural Net- works (CNNs) are known to have a good performance on data with high locality when words get more care weight about the features surrounding them. For our classification problem, we are trying to get high locality in the text given their short length and their tendency to focus on cyberbullying. We implemented CNNs that received input text in the form of sequences of integer representations of stemmed unigrams. Our character processing involved the conversion of emoticons into word representations and the removal of non-Latin characters. We also removed frequently occurring URL components (e.g., names of popular websites), metadata encoded in the main body text (e.g., 'RT: '), and a variety of social media platform- specific features. Hashtags and @-mentions were reduced to binary features. The text was then lower-cased and tokenized using NLTK's TweetTokenizer3. The tokenized text was next encoded utilizing a dictionary of integers, with the original ordering of the tokens preserved.

Long short Term Memory networks, LSTMs have been observed as the most useful solution. LSTMs have an edge over traditional feed-forward neural networks and RNN in many ways. This is because of their property of selectively memorizing patterns for long durations of time. LSTMs, make small changes to the information by multiplications and additions. With

LSTMs, the data flows through a mechanism known as cell states. This way, LSTMs can selectively recall or forget things. The data at a particular cell state has three dependencies. It is used for processing, predicting and classi- fying based on time series data.

A typical LSTM network is comprised of different memory blocks called cells which are responsible for remembering things and manipulations to this memory is done through three major mechanisms, called gates.

a) Forget gate The information that's not now useful in the cell state is removed with the forget gate. Two inputs are fed to the gate and multiplied with the weight matrices proceeded by the addition of bias. The resultant is passed through an activation function which gives a binary output. If for a particular cell state the output is 0, the information is forgotten and for the output 1, the information is used for the future use.

b) Input gate the addition of useful information to the cell state is done by input gate. First, the information is regulated using the sigmoid function and filters the values to be remembered similar to the forget gate using the two inputs. Then, a vector is created using tanh function that gives output from -1 to 1. Atlast, the values of the vector and the regulated values are multiplied to obtain the useful information.

c) Output gate the task of extracting useful information from the current cell state which is to be presented as an output is done by output gate. First, a vector is generated by applying the function tanh on the cell. Then, the information is regulated using the sigmoid function and filters the values to be remembered using the two inputs. Atlast, the values of the vector and the regulated values are multiplied which are then sent as an output and input to the next cell.

6. RESULT

The model will generate the result based on the tweets provided as bullying or non-bullying tweet. The confusion

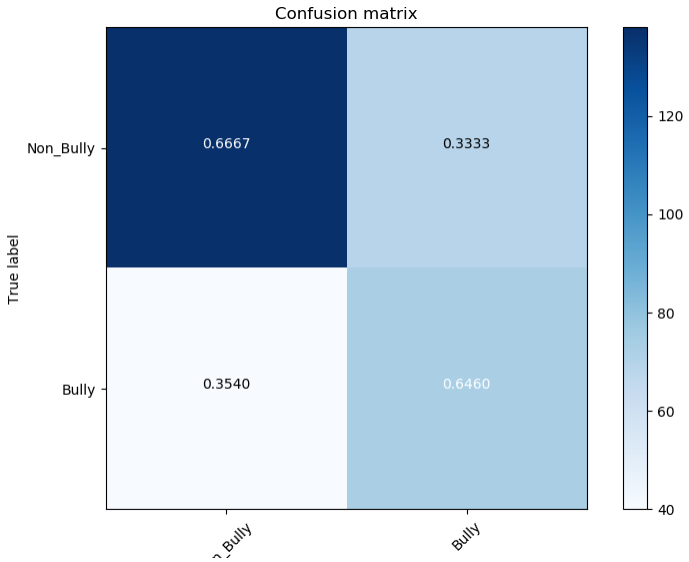

Fig. 2. Confusion Matrix

Fig. 3. Output for non-bullying

matrix in Fig 2 has been visualised for the bullying and non-bullying tweets. It provides the accuracy of the cyber bullying detection model. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made. Our study shows that the DNN models were adaptable and transferable to the new dataset. DNN based models coupled with transfer learning outperformed all the previous results for the detection of cyberbullying in this Titter dataset using Deep Learning models. Fig 3 illustrates the output for non-bullying tweet. Fig 4 illustrates the output for bullying tweet.

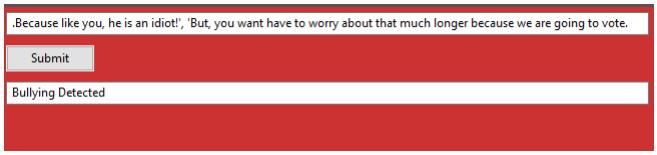

Fig. 4. Output for bullying

7. CONCLUSION

Technology revolution advanced the quality of life, how- ever, it gave harassers a solid ground to conduct their harmful crimes. Internet crimes have become very dangerous since victims are targeted everytime and there are no chances for escape. Cyberbullying is one of the most critical internet crimes and research proved its critical consequences on victims. From suicide to lowering victims self-esteem, cyberbullying control has been the focus of many psychological and technical research. In this paper, the issue of cyberbullying detection on Twitter has been tackled. The aim was to advance the current state of cyberbullying detection by shedding light on critical problems that have not been solved yet. To the best of our knowledge, there has been no research that considered eliminating features from the detection process and automating the process with a CNN. The proposed algorithm makes cyberbullying detection a fully automated process with no human expertise or involvement while guaranteeing better result. Comprehensive experiments proved that deep learning models outperformed classical machine learning approaches in cyberbullying problem.

8. ACKNOWLEDGMENT

We would like to express our sincere gratitude and thanks to our project guide Prof. Kumud Wasnik, as she has been both the guiding light for this project and has provided us with the best knowledge, advice and encouragement which helped us in the successful completion of this project. We express our sincere thanks to our principal Dr. Sanjay Pawar for his constant moral support. He has rightfully provided us insight and expertise that assisted the research. We are also thankful to our colleagues and teachers for their help in offering us the resources in running the project.

REFERENCES

1. Haipeng Zeng, Hammad Haleem, Xavier Plantaz, NanCao and Huamin Qu “CNN Comparator: Comparative Analytics of CNN” arXiv, 15 Oct,2017.

2. Rui Zhao, Kezhi Mao “CyberBullying Detection based on Semantic - Enhanced Marginalized Denoising Auto-encoders” IEEE Transaction on Affective Computing, 2015.

3. A Simple way to Prevent Neural Networks from Overfitting” Journal of Machine Learning Research 1929-1958,2015

4. Alexis Conneau, Holger Schwenk, Yann Le cun “Very Deep CNN for Text Classification” Association for Computational Linguistics, Volume1, pages 1107-1116,7 April 2017.

5. Elaheh Raisis,Bert Huang “Cyberbullying Detection with Weakly Su-pervised Machine Learning” International Conference on Advances in Social Networks Analysis and Mining IEEE/ACM,2017.

6. Alexis Conneau, Holger Schwenk, Yann Le cun "Very Deep CNN for Text Classification" Association for Computational Linguistics, Volume1, pages 1107-1116,7 April 2017

7. Haipeng Zeng,Hammad Haleem,Xavier Plantaz,NanCao and Huamin Qu "CNN Comparator: Comparative Analytics of CNN" arXiv,15 Oct,2017.

8. Nitish Srivastava, Geoffrey Honton,Alex Krizhevsky,Iiya Sutskever,Ruslan Salakhutdinov "Dropout: A Simple way to Prevent Neural Networks from Overfitting" Journal of Machine Learning Research 1929-1958,2015

9. Van Hee, Cynthia, et al. "Automatic Detection of Cyberbullying in Social Media Text." PloS One, Public Library of Science, 8 Oct. 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC6175271/.

10. Prabhu. "Understanding of Convolutional Neural Network (CNN) - Deep Learning." Medium, Medium, 21 Nov. 2019, medium.com/@RaghavPrabhu/understanding-of-convolutional-neural- network-cnn-deep-learning-99760835f148.

11. Cao, et al. "An Improved Convolutional Neural Network Algo- rithm and Its Application in Multilabel Image Labeling." Com- putational Intelligence and Neuroscience, Hindawi, 4 July 2019, www.hindawi.com/journals/cin/2019/2060796/.

12. "Introduction to Convolution Neural Network." GeeksforGeeks, 21 Aug. 2017, www.geeksforgeeks.org/introduction-convolution-neural- network/.

13. Brownlee, Jason. "A Gentle Introduction to Long Short-Term Mem- ory Networks by the Experts." Machine Learning Mastery, 19 Feb. 2020, machinelearningmastery.com/gentle-introduction-long-short-term- memory-networks-experts/.