Volume: 9 Issue: III Month of publication: March 2021

DOI: https://doi.org/10.22214/ijraset.2021.33229

Implementation of Cyberbullying Detection using Machine Learning Techniques

Saloni Kargutkar[1], Vidya Chitre[2]
[1]PG Scholar, [2]Professor, Department of Information Technology, Vidyalankar Institute of Technology, Wadala, Mumbai

Abstract: Cyberbullying could be an upsetting on-line wrongdoing with disturbing consequences. It seems in several forms, and in most of the social networks, it's in text format. Automatic detection of such incidents needs intelligent systems. Most of the prevailing studies have approached this drawback with typical machine learning models and therefore the majority of the developed models in these studies are applicable to one social network at a time. In recent studies, deep learning primarily based models have found their means within the detection of cyberbullying incidents, claiming that they will overcome the restrictions of the standard models, and improve the detection performance. Cyberbullying is that the use of technology as a medium to bully somebody. Though it's been a difficulty for several years, the popularity of its impact on teenagers has recently inflated. Social networking sites offer a fertile medium for bullies, and youths and young adults UN agency use these sites are susceptible to attacks. Through machine learning, we are going to realize language patterns utilised by bullies and their victims, and develop rules to automatically realize cyberbullying content. We find that our approach is with success able to determine vital variations between cyberbullying and regular media sessions, and supply a performance increase in cyberbullying detection. This paves the means for a lot of nuanced work on the utilization of temporal modelling to find and mitigate the incidence of cyberbullying.

Keywords: Cyberbullying, Deep learning, Machine learning, Content based cybercrime, Social Networking

I. INTRODUCTION

Day to day life is entirely or in a way dependent on the advent of internet. Thus, cyberbullying has been a major worry. With the advancement in technology, the internet has been a safe and secure sphere of communication, though the arena of social media has been prone to cybercrimes. Since the social lifestyle surpass the physical barrier of human interaction and affords inappropriate interaction with unknown people, it is important to analyse and study the domain of cyberbullying. Moreover, a well-specified law framework for cyberbullying has not been enforced in majority of the countries, therefore the data to defend the matter is unsure. Cyberbullying are often outlined because the use of a web communication to bully or harass an individual World Health Organization doesn't have potential to react [1], usually by causation messages of a threating or discouraging nature. It is evident that around 87 percent of the today's youth have witnessed some form of cyberbullying [2]. One of the rare reports [3] on cyberbullying states that 60 percent of Gulf Countries' youth overtly admit the presence of cyberbullying amongst their peers. This study conjointly states that solely quarter the predators on-line do bully their victims offline. This means that net have impressed three quarters of the predators to bully others, whereas they wouldn't have thought-about bullying physically. Cyberbullying can take various forms like Sexual Harassment, Hostile Environment, Racism, Revenge, and Retaliation. Since the offender is hidden to the victim, the problem statement gets complex. Effects of cyberbullying can range from temporary anxiety to suicide [6]. This is the reason cyberbullying is an interesting field of research. The adverse result of a cybercrime are often drastic- Cyberbullying was powerfully connected dangerous thinking compared with ancient bullying (JAMA pediatrics, 2014) [4], Hence, the requirement for a good system to spot cyberbullying and relieve the plight of distressed users. Since, cyberbullying can take place without the direct confrontation of the perpetrator, it is lot more vulnerable. Moreover, the most vicious state of bullying is that it can take place across social networks which were previously unreachable. Thus, with the proliferation of social media and web access, the act of cyberbullying too has inflated manifold. Twitter is one amongst the foremost lauded and widespread social media existing. It allows users to send and browse 140-character message. It is astonishing that about 330 million active users access the platform and nearly 500 million tweets are exchanged a day. Since about 80 percent of the user's access with their mobile phones, it has been an arena of real-time communication. A study determined that Twitter is turning into a cyberbullying playground (Xuetal, 2012). In this analysis, we tend to tend to utilize this important information and knowledge within the kind of tweets to enhance the prevailing cyberbullying detection performance. Since, Twitter is very user-friendly it enables the use of extended features like network, activity, user and tweet content, to train our detection model and improve its performance.

A Convolution Neural Network (CNN) popularly known as ConvNet is a specific type of artificial neural network that use perceptron's, a machine learning algorithm to analyse data. CNNs apply to natural language processing, image processing and other cognitive tasks. [10] [11] [12]. Our novel idea is 2 to implement the features of CNN used for image analysis and process the same for text analysis. Since, text can be defined as an arrangement of pixels in an organized way, the method of CNN is effectively used for calculation. [10]

II. RELATED WORK

Cyberbullying is known as an occurrence at least since 2003 [5]. Use of social media exploded with launching of multiple platforms such as Facebook (2004), Orkut (2004), MySpace (2003), Wikipedia (2001), and Twitter (2005). By 2006, researchers had pointed that cyberbullying was as serious phenomenon as offline bullying [7]. However, automatic detection of cyberbullying was addressed only since 2009 [8].

Cyberbullying detection is a text classification problem, as a research topic. Most of the existing works fit in the following template: get training dataset from single SMP, engineer variety of features with certain style of cyberbullying as the target, apply a few traditional machine learning methods, and evaluate success in terms of measures such as accuracy and F1 score. These works heavily depends on handcrafted features such as use of swear words.

These ways is perhaps aiming to own low preciseness for cyberbullying detection as handcrafted features are not robust against variations in bullying style across SMPs and bullying topics. Deep learning has been applied for cyberbullying detection recently [9]. Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA) developed via semantic extension of the popular deep learning model stacked denoising auto encoder is a new representation learning method used in a research topic. [13]. In this research [14], a weakly supervised machine learning technique for all the while surmising user rolls in provocation based bullying and new vocabulary markers of harassment.

P. Zhou, et. al. [15] proposed attention based B-LSTM technique, this can automatically concentrate on the words that have conclusive impact on classification, to catch the most imperative semantic information in a sentence, without utilizing additional knowledge and NLP frameworks.

It was analysed in this research [16] cyberbullying corpuses using the bag-of-words model to find the most common used terms by cyberbullies and used them to create queries capable of reaching a precision of 91.25% on average.

III.IMPLEMENTATION

The proposed system is implemented in Python and TensorFlow. TensorFlow is a high-performance computing framework which is widely used in research, development and analysis in the fields of data science and deep learning. The Twitter dataset used consists of 69874 tweets, which are converted to vectors using open source word embedding Glove. These messages were sorted and labels were generated. Neural Network model revealed here were implemented utilizing Keras on top of TensorFlow. We pre-process the data, exposing it to standard tasks of expulsion of stop words, accentuation marks and lowercasing, before clarifying it to allocating individual labels to each remark.

A. Message feed

The dataset feeds messages that is needed by the input system. This is often the input for word embedding and starting of the advancement within the system.

B. Word Embedding

It is capable of capturing context of a word, semantic and syntactic similarity, relation with alternative words, etc. The data from the message feed is embedded into numerical type for the input of the CNN. Every word is described by a true worth vector. The distributed illustration of words is learned by the technique of transfer learning.

C. Deep Neural Network Models (DNN)

All the models have similar layers apart from the neural architecture layer that is exclusive to every model. The embedding layer, which can be explained in additional detail within the following section, processes a fixed-length sequence of words. Then there's the absolutely connected layer that may be a dense output layer with the abundance of neurons adequate to the amount of categories. The outer layer is that the softmax layer that provides softmax activation. Each model is trained exploitation backpropagation with Adam optimizer and categorical cross entropy loss perform.

1) *Convolution Neural Network:* In our research, the notion of CNN implementation is included. Convolutional Neural Networks Convolutional Neural Net- works (CNNs) square measure proverbial to own a decent performance on knowledge with top locality once words get additional care weight regarding the options close them. For our classification downside, we have a tendency to try to urge high locality within the text given their short length and their tendency to target cyberbullying. CNN is used with multiple layers which provide a process of iterative analysis over different layers to provide an efficient and accurate analysis. Hence, a large corpus of tweets is obtained using twitter APIs, which is undergone a series of data pre-processing to provide a clean dataset which is then trained and tested. Vectors generated by word embedding is that the input for the neural network layers. The first layer is the Convolution layer of CNN output of this layer is given to the max-pooling layer and fully connected network is generated. Softmax function is employed once the absolutely connected layer to get the output.

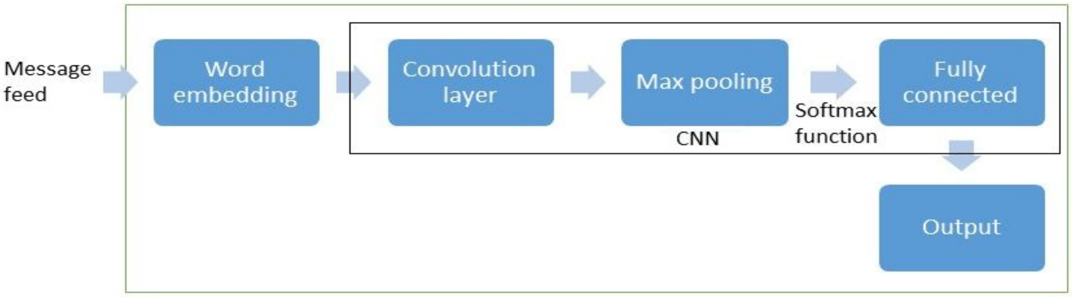


Fig. 1 System Flow

Convolutional Neural Networks (CNNs) square measure better-known to possess smart performance on information with high neighbourhood, once words get a lot of care weight concerning the options close them. We have a tendency to used CNNs that received input text within the variety of sequences of whole number representations of stemmed unigrams. The text was then lower-cased and tokenized victimisation NLTK's TweetTokenizer3. The tokenized text was next encoded employing a wordbook of integers, with the first ordering of the tokens preserved. The encoded text was converted into dense vectors of fixed size. The kernels were optimized using TensorFlow Adam optimizer () using categorical cross-entropy as the loss function. The output dimensions were flattened using a sigmoid function into two output nodes whose values are floats between 0 and 1, with 1 representing positive and 0 representing negative. To test the performance of our model, we took 70% of the dataset as training set, and 30% of it for testing.

2) *Long Short-Term Memory(LSTM):* Long short Term Memory networks, LSTMs have been noticed as the most useful solution. LSTMs have an advantage over traditional feed-forward neural networks and RNN in numerous ways. This is owing to their property of by selection memorizing patterns for long durations of your time. LSTMs, make small changes to the information by additions and multiplications. With LSTMs, the data flows through a mechanism known as cell states. This way, LSTMs can selectively forget or recall things. There are three dependencies for the data at a particular cell state. It is used for process, predicting and classifying supported statistic information. A typical LSTM network is comprised of various memory blocks known as cells that memory blocks accountable for basic cognitive process things and manipulations to the present memory is completed through 3 major mechanisms, known as gates.

a) *Forget Gate:* The information that's not currently helpful within the cell state is removed with the forget gate. Two inputs are fed to the gate and increased with the burden matrices proceeded by the addition of bias. The resultant is moved through an activation function which produces a binary output. If for a specific cell state the output is 0, the information is forgotten and for the output 1, the knowledge is employed for the longer term use.

b) *Input Gate:* The addition of beneficial information to the cell state is done by input gate. First, the information is managed using the sigmoid function and clarify the values to be remembered similar to the forget gate using the two inputs. Then, a vector is formed victimization tanh perform that provides output from -1 to one. At last, the values of the vector and also the regulated values area unit increased to get the helpful info.

c) *Output Gate:* The task of extracting helpful data from this cell that is to be introduced as an output is finished by output gate. Foremost, a vector is generated by applying the function tanh on the cell. Then, the information is managed using the sigmoid function and filters the values to be remembered using the two inputs. Finally, the regulated values and values of the vector are multiplied which are then sent as an output and input to the next cell.

The LSTM consists of units or memory blocks within the continual hidden layer that contains memory cells with self-connections storing the temporal state of the network. Additionally to the current the network has special increasing units referred to as gates to manage the flow of knowledge within the network. Every memory block within the original design contained associate input gate associated an output gate. The input gate controls the flow of input activations into the memory cell and also the output gate controls the output flow of cell activations into the remainder of the network. The forget gate was accessorial to the memory block that scales the inner state of the cell before adding it as input to the cell through the self-recurrent affiliation of the cell, so adaptively forgetting or resetting the cell's memory. Additionally, the fashionable LSTM design may also contains spyhole connections from its internal cells to the gates within the same cell that facilitate to be told precise temporal arrangement of the outputs.

IV. RESULT

The results and discussion deals with the analysis of comparison graph on the basis of precision and run time complexity. When the algorithm runs it takes each twitter messages and breaks it down into individual words. Each word is compared to the words in the bully dictionary. If it matches any of the words then it is added to the precision value. The precision values are compared with the threshold value and if it results in less than the threshold value it is reviewed as cyberbullied message. In the end, the algorithm adds all the twitter messages having precision values less than the threshold.

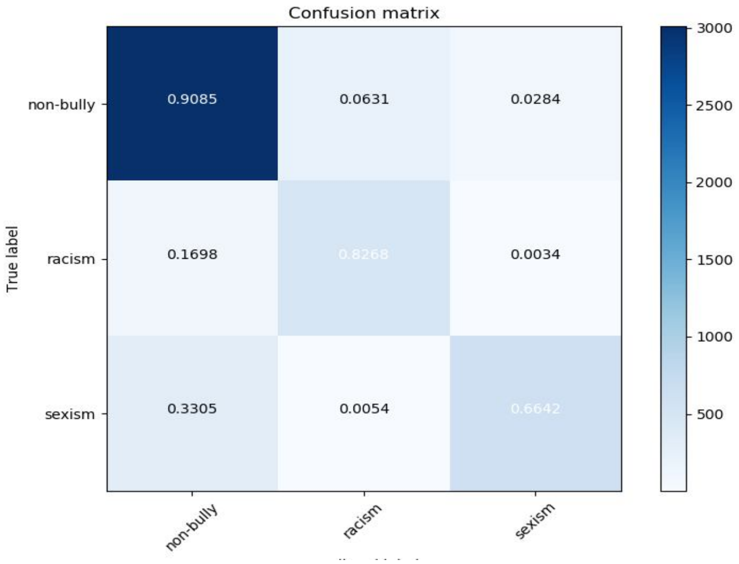
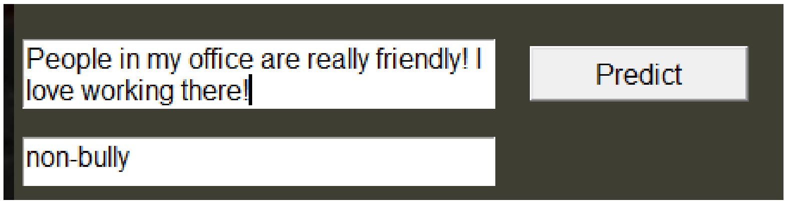


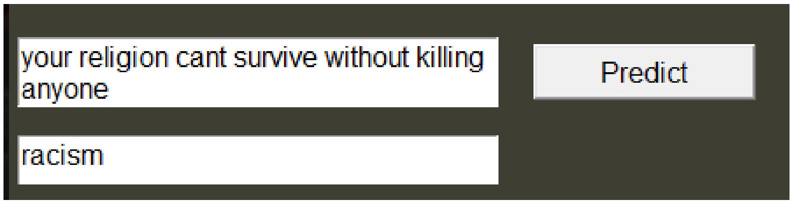
Fig. 2 Confusion Matrix



People in my office are really friendly! I love working there!

non-bully

Fig. 3 Output for non-bullying



your religion cant survive without killing anyone

racism

Fig. 4 Output for bullying

The model will produce the result based on the tweets provided as non-bullying or bullying tweet. The confusion matrix in Fig 2 has been visualised for the bullying and non-bullying tweets. It delivers the precision of the cyber bullying detection model. A confusion matrix is an abstract of prediction results on a classification problem. The number of incorrect and correct predictions are summarized with count values and then broken down by each class. This is the solution to the confusion matrix. The confusion matrix shows the ways in which the classification model is confused when it produces predictions. It provides us insight for not only into the errors being made by a classifier but also more importantly the types of errors that are being made. Our research shows that the DNN models were adaptable and transferable to the new dataset. DNN based models coupled with transfer learning surpassed all the previous results for the detection of cyberbullying in this Twitter dataset using Deep Learning models. The Output for non-bullying tweet is represented in Fig 3. Fig 4 represents the output for bullying tweet.

V. CONCLUSION AND FUTURE SCOPE

In this study, we successfully implemented detection of cyberbullying incidents in social media platforms using DNN based models. We further expanded our work by using a social media dataset, Twitter, to investigate the adaptability and transferability of the models to the new dataset and also to compare the performance of the DNN models against the conventional ML models which were used in previous studies on the Twitter dataset for cyberbullying detection. The datasets for cyberbullying detections contains very few posts marked as bullying. This imbalance problem can be partially covered by oversampling the bullying posts. However, the effects of such prevalence on the performance of models need to be further evaluated. Our research shows that the DNN models were transferable and adaptable to the new dataset. DNN based models associated with transfer learning surpassed all the previous results for the detection of cyberbullying using ML models in this Twitter dataset.

Although many researches addressed cyberbullying in SMP (social media platform), the techniques used for the detection proves inefficient in classification. In this system, we represent a new approach for the detection of cyberbullying. This system uses convolution neural network algorithm which operates through many layers and gives accurate classification. Thus a more intelligent way, compared to the traditional classification algorithms is designed. Future scope includes enhancing the effectiveness of the word-pronunciation conversion and connect pronunciation features with the CNN model more tightly.

REFERENCES

[1] R. Shetgiri, "Bullying and Victimization among Children", Advances in Pediatrics, vol. 60, no. 1, pp. 33-51, 2013.

[2] K. Poels, A. DeSmet, K. Van Cleemput, S. Bastiaensens, H. Vandebosch and I. De Bourdeaudhuij, "Cyberbullying on social network sites. An experimental study into bystanders," Cyberbullying on social network sites, vol. 31, p. 259–271, 2014.

[3] ICDL, "Cyber Safety Report: Research into the online behaviour of Arab youth and the risks they face," ICDL Arabia, 2015.

[4] "Text classification using convolution neural networks.Glenn Sterner, Department of Sociology and Criminology, The Pennsylvania State University, University Park, PA, USA Diane Felmlee, Department of Sociology and Criminology, The Pennsylvania State University, University Park, PA, USA, 2017.

[5] R. L. Servance. Cyberbullying, cyber-harassment, and the conflict between schools and the first amendment. Wisconsin Law Review, pages 12–13, 2003

[6] S. Hinduja and J. W. Patchin. Bullying, cyberbullying, and suicide. Archives of suicide research, 14(3):206–221, 2010.

[7] J. W. Patchin and S. Hinduja. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. Youth violence and juvenile justice, 4(2):148–169, 2006.

[8] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. Detection of harassment on web 2.0. In The workshop on Content Analysis in the WEB 2.0, WWW, pages 1–7, 2009.

[9] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. In WWW, pages 759–760, 2017.

[10] Johnson, Rie, and Tong Zhang. "Effective use of word order for text categorization with convolutional neural networks." arXiv preprint arXiv: 1412.1058 (2014).

[11] Zhang, Xiang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification." Advances in neural information processing systems. 2015.

[12] A. J. McMinn, Y. Moshfeghi and J. M. Jose, "Building a large-scale corpus for evaluating event detection on twitter," in 22nd ACM international conference on Information & Knowledge Management, 2013.

[13] Rui Zhao, Kezhi Mao "CyberBullying Detection based on Semantic-Enhance Marginalize Denoising Autoencoders" IEEE Transaction on Affective Computing, 2015.

[14] Elaheh Raisi, Bert Huang "Weakly Supervised Cyberbullying Detection with Participant Vocabulary Consistency" Social Network Analysis and Mining, May 24, 2018.

[15] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Houng Wei, Hao,Bo Xu"Attention-based Bi-directional Long Short Term Memory Network for Relation Classification" proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,pages 207- 212,August 12,2016.

[16] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: query terms and techniques," in Proceedings of the 5th annual acm web science conference, pp. 195-204, 2013.