Cyber Bullying Detection Based on Twitter Dataset

Chapter · October 2020

DOI: 10.1007/978-981-15-7106-0_9

3 authors, including:

Debajyoti Mukhopadhyay
Bennett University
251 PUBLICATIONS **1,553** CITATIONS

SEE PROFILE

Cyber Bullying Detection Based on Twitter Dataset

Debajyoti Mukhopadhyay, Kirti Mishra, Kriti Mishra, and Laxmi Tiwari

Abstract The acceleration of different social media platforms has alternated the way people communicate with each other it has also ensued in the rise of Cyber-bullying cases on social media that has various adverse effects on an individual's health. In this project, we aim to build a system that tackles Cyber bully by identifying the mean-spirited comments and also categorizing the comments into peculiar division. The target of developing such a system is to deal with Cyber bullying that has become a prevalent occurrence on various social media. The system uses two noticeable features—Convolutional Neural Network and Long Short-Term Memory which improves the efficiency of the system.

Keywords Cyber bullying detection · CNN algorithm · Twitter · Social media harassment · Online harassment · Long short-term memory · Word embedding

1 Introduction

Social media is the use of virtual platform for connecting, interacting, sharing of contents and opinion around the globe. Since the development of social platform, its usage by teens and adults across the globe has seen great upsurge. The most famous

D. Mukhopadhyay (✉) · K. Mishra · K. Mishra · L. Tiwari
Computer Science Department, Mumbai University, Mumbai, Maharashtra, India
e-mail: debajyoti.mukhopadhyay@gmail.com

K. Mishra
e-mail: mishrakirti2403@gmail.com

K. Mishra
e-mail: kritismishra41@gmail.com

L. Tiwari
e-mail: laxmitiwari21998@gmail.com

D. Mukhopadhyay
WIDiCoReL Research Lab, Mumbai, Maharashtra, India

A. Joshi et al. (eds.), *Machine Learning for Predictive Analysis*,
Lecture Notes in Networks and Systems 141,
https://doi.org/10.1007/978-981-15-7106-0_9

social media platforms are Facebook, YouTube, Twitter, Instagram. The paramountcy of it can be implied by the fact that out of 7.7 billion people in the world at least 3.5 billion people are online. Social media had a vital role in connecting people and eventually made the world a smaller well-connected and more tolerant place. But let's not blow the fact that social media is a double-edged sword and can have a various detrimental effects which can be of serious ramification like circulation of rampant information to cyber bullying.

Cyber bullying is a kind of bullying that occurs over digital devices that include phones, laptops, computers, tablets, netbook, hybrid through various SMS, apps, forums, gaming which are intended to hurt, humiliate, harass and induce various negative emotional responses to the victim, using text, images or videos and audios. It can cause more suffering than traditional bullying as the atrocious messages are perpetual and easy to prey on potential victims. Cyber bullying can result in low self-esteem in victims as constant mean messages can result in victims being more anxious and insecure about themselves. It may result in poor performance in school grades among teenagers. Teenagers and kids who are unable to cope up with bullying may result in social isolation by skipping school and interaction with friends and family, and also indulge in activities like drugs and alcohol. It also affects adults in prosaic day to day life activities. Victims of cyber bullying can have physical effects like headache, stomach problems and issues which are created due to stress like various conditions of skin and also ulcers of stomach. Victims may have eating disorders and various weight-related issues and sleeping disorders like insomnia. Apart from the physical effects, it may also have psychological effects like anger, frustration, sadness, behavioural problem like losing interest and in adverse condition may result in suicide intention. Cyber bullying has the ability to take the whole world into its grip by spreading false information regarding politics, diseases, laws and many more. Targeting a person based on appearance, different ideology, colour, chauvinism, sexual preference is a familiar occurrence. It often passes prejudiced and hatred towards targeted person or group. Cyber bullying has created a lot of hue and cry in the world and has created a compelling situation that has to be dealt with by recognizing such activities instantaneously and to develop stringent laws protecting people against search felony [1–8].

From Fig. 1, it can be observed that most cases of cyber bullying have been proclaimed by India. According to research conducted by Symantec, it is estimated that out of every 10 individuals nearly 8 individuals have been victims of Cyber bullying of some form in the nation accompanied by Brazil and the United States. Out of all the social media platforms, Twitter is one such means by which bullying occurs. It is an indispensable way for considerable socializing and to affix with like-minded people to prorate opinion. It also endows a platform to monitor brands and its eminence while keeping up with the voguish news around the globe and has seen a gradual rise in cyber bully associated cases. Hence, in this system, we have used Twitter dataset which focuses on bullying related to text.

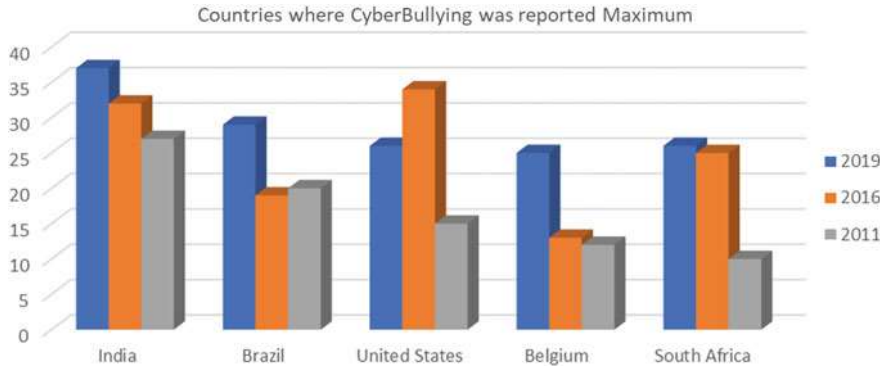

Fig. 1 Graphical representation of countries where Cyber bullying was reported maximum

2 Proposed Approach

The proposed approach using Convolutional Neural Network (CNN) algorithm for Cyber bullying Detection aims to improve the efficiency of the system that used long feature extraction and selection methods resulting in tedious and time-consuming process. CNN algorithm targets to improve the classification of tweets categorizing it into bullying or non-bullying divisions by using a word embedding an approach where words with analogous meaning have similar representation in the form of vectors which efficaciously saves the process of feature determination and extraction. As in the process of Feature determination and extraction method, the features can be entered manually or automatically that are considered to be relevant to the matter dealt in a text, in this case, bullying. With the exponential number of tweets abounding features are added which only makes things more conglomerate. It is then passed to the classifier and thus the use of word embedding saves the effort. For training of the system labelled data is used. After the training period, the System will detect cyber bully related tweets that have matching keywords from the trained database. In order to determine the cyber bully, we are focusing on the keywords posted by the users.

3 Methods

Figure 2 represents an overview of the proposed system. Initially, it starts with fetching of tweets from Twitter database or we can enter the tweets as an input using the keyboard too. Tweets are fetched and then pre-processed using a python code such that stop words, noisy and irrelevant data are removed and the processed words are then tokenized. A matrix-vector is formed to capture semantics and model words using word embedding (Fig. 3).

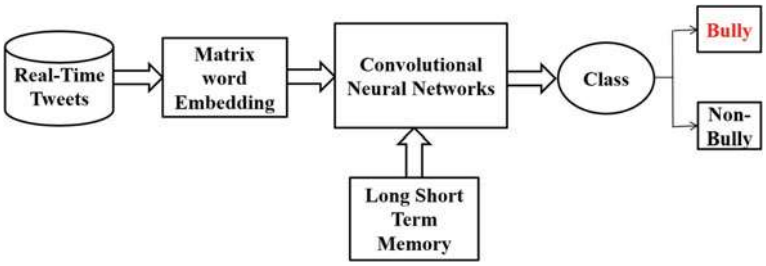

Fig. 2 Overview of the proposed approach

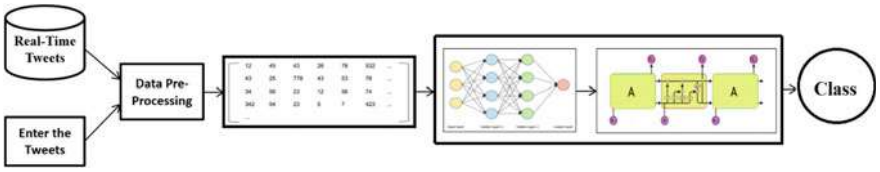

Fig. 3 Overall architecture

3.1 Data Pre-processing

Data pre-processing is a technique by which raw data is converted into useful information. In the data cleaning stage, we remove punctuation, emojis and removing of a numerical along with the lowercase conversion takes place. It also eliminates noisy and irrelevant data in the next stage tokenization take place in which sentence of the tweets are divided into certain characters, for example, the sentence "please give me three apples" would be divided into word "please, give, me, three, apples". The stop words are removed after the tokenization the stop words are words like "be, can, is, the" which can be ignored there is no standard list for stop words. In the last step of data pre-processing, we reduced the inflected words, and hence words like happily, happiness and happy are reduced to happy as the word happy is lemma of all these words (Figs. 4 and 5).

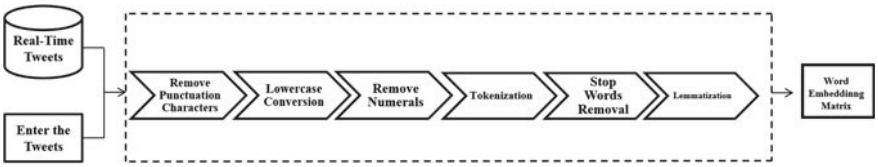

Fig. 4 Steps involved in data pre-processing

Fig. 5 Representation of data pre-processing

3.2 Convolutional Neural Network

CNN algorithm compartmentalizes tweets in more intelligent methodology than any other classification algorithm. CNN also jettisons the work required in the traditional cyber bullying detection system by adapting major principles instead of long machine learning classification techniques. This method of implementation also eliminates the few layers that were used in other traditional classification algorithms and acts as a remarkable aspect. Convolutional neural networks are inclusive of neurons called Covnets. Covnets in CNN share their parameters and is designed to meet the needs for classification. CNN algorithm is a multilayer perceptron and useful in natural language processing. The fact that they are useful for these fast-roaring growth areas is one of the main reasons they are so important in Deep Learning and Artificial Intelligence technology. Each of these layers contains neurons that are connected to neurons in the previous layer.

3.2.1 Input Layers

Number of neurons in input layer is equal to the total number of features in our data. It accepts input in different forms.

3.2.2 Hidden Layers

Input from the previous layer (input layer) is forwarded to further neurons that are present in between input and output layer and are called Hidden Layers. In these layers, number of neurons is greater than number of features. It performs calculations on the input.

3.2.3 Output Layers

Output of hidden layer is fed into logistic function which converts output of each class into probability score of each class. It delivers the outcome of the calculations and extractions. Each neutron has its own weight. Instead of neurons being connected to every neuron in pre-layer, they are instead only connected to neurons close to it and all have the same weight. This simplification in networks means the new network upholds the spatial aspect of the data set.

3.3 Long Short-Term Memory

A long short-term memory is networks which is a type of recurrent neural network and are capable of retaining the information for a long period of time. A recurrent neural network which is used for creating memory in neural network has a hidden state which helps to store information about the past. It also allows us to update the hidden state and solves the problem of sequence prediction partially. But fails because it is possible that the relevant information is not present at the site where it is needed because of many irrelevant data and it cannot predict the words stored in long memory but instead is used for recent information, for example, an apple has colour _____. The answer anticipated would be red. As in this context, the RNN has relevant information required to make prediction. Furthermore, if there is a statement I bought apples from the market. They are very delicious and its colour is _____ the network needs context apple from the previous statement and it is possible that the gap between two statements is more and the network will not be able to associate the information. And hence LSTM is used which has a property of selectively remembering patterns with the help of different memory blocks called cells. The ability to remember is not learnt but a default nature of the LSTM. The memory blocks are managed using the gates.

3.3.1 Forget Gate

The inputs which are given to the network are multiplied with the matrix and the addition of bias is done. It is then passed through activation function which gives output 0 or 1. If the output is zero then the pattern is not remembered similarly if the output is 1 then the pattern is remembered.

3.3.2 Input Gate

All the useful information in the network is passed using input gate.

Fig. 6 Output: Bullying detected

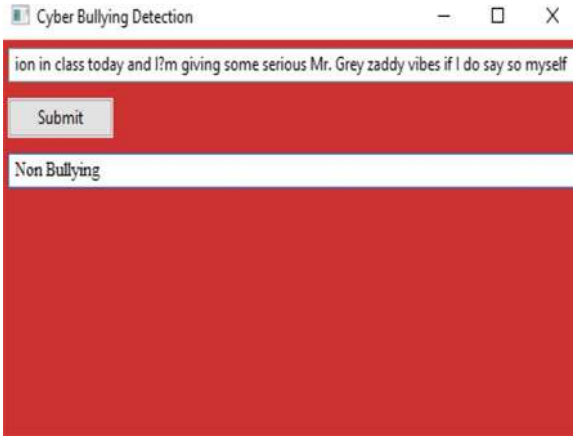

Fig. 7 Output: Non bullying

3.3.3 Output Gate

It determines the output to be generated.

4 Results

Figures 6 and 7 symbolize the end product of the proposed methodology where the system effectually distinguishes the text into its relevant category.

5 Future Work

Several possible optimizations for future works are as follows:

1. As sending image and video is becoming popular among adolescents. Hence, image and video processing would be another important area for cyber bullying detection.
2. Cyber bullying can also be further improved to take a variety of actions depending on the perceived seriousness of the post.
3. Detecting cyber bullying in streaming data.
4. Evaluating Annotation judgement.

6 Conclusion

This paper proposes a system to detect cyber bullying in real-time by using Twitter API. As social media is an emerging platform to connect worldwide and easy source to attack anyone in many forms of danger like cyber bullying. Automatic detection of cyber bully would enhance moderation and allow to respond quickly when necessary including different types of cyber bully covering posts from bullies and victims. We, therefore, intend to apply deep learning techniques to improve classifier performance.

References

1. R. Zhao, K. Mao, Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoders. IEEE Trans. Affect. Comput. (2015)
2. V. Nandakumar, B.C. Kovoor, MU Sreeja, Cyber-bullying revelation in twitter data using Naive-Bayes classifier algorithm. Int. J. Adv. Res. Comput. Sci. **9** (2018)
3. S. Bhoir, T. Ghorpade, V. Mane, Comparative analysis of different word embedding models (IEEE, 2017)
4. E. Raisis, B. Huang, Cyberbullying detection with weakly supervised machine learning, in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (2017)
5. E. Raisi, B. Huang, Weakly supervised cyberbullying detection with participant vocabulary consistency. Soc. Netw. Anal. Min. (2018)
6. P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Wei, B. Xu, Attention-based bi-directional long short-term memory network for relation classification, in *Proceedings of the 54th Annual Meeting of the Association For Computational Linguistics*, 12 Aug 2016, pp. 207–212
7. A. Conneau, H. Schwenk, Y.L. Cun, Very deep CNN for text classification, vol. 1. Association for Computational Linguistics, pp. 1107–1116 (2017)
8. https://www.comparitech.com/internet-providers/cyberbullying-statistics/