ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 12, December 2018

Classification of Cyber-bullying using Convolution Neural Network

Koushik Sai Venkataramanan

Department of Information Technology, SRM Institute of Science & Technology, Tamil Nadu, India

Abstract: More than 1.96 billion are bound to have an inevitable social life. However, the growing decade poses serious challenges and the online-behaviour of users have been put to question. Increasing cases of harassment and bullying along with cases of fatality have been a serious issue. Though, many old-school models are available to control the mishap, the need to effectively classify the bullying is still feeble. To effectively monitor the bullying in the virtual space and to stop the deadly aftermath with implementation of Machine Learning and Language processing. In this paper, we propose a methodology to provide a binary classification of cyberbullying. Our method uses an innovative concept of CNN for text analysis however the existing methods use a naive approach to provide the solution with less accuracy. An existing twitter dataset is used for experimentation and our framework is verified with other existing procedures and is found to provide better accuracy and classification.

Keywords: Convolution Neural Network (CNN), Cyberbullying, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA)

1. INTRODUCTION

The advent of internet show manifold of users with their life entirely or in a way dependent on it. 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 implemented in majority of the countries, thus the knowledge to defend the problem is uncertain. Cyberbullying can be defined as the use of electronic communication to bully a person, typically by sending messages of an intimidating or threating nature. It is evident that around 87 percent of the today's youth have witnessed some form of cyberbullying.

Cyberbullying can take various forms like Sexual Harassment, Hostile Environment, Revenge, and Retaliation. Since the offender is hidden to the victim, the problem statement gets complex. This is the reason cyberbullying is an interesting field of research. The adverse effect of a cybercrime can be drastic- Cyberbullying was strongly related suicidal ideation in comparison with traditional bullying (JAMA Paediatrics, 2014) [1]. Hence, the need for an effective system to identify 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 internet access, the act of cyberbullying too has increased manifold. Twitter is one of the most lauded and popular social media existing. It enables users to send and read 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 research, we tend to utilize this vital data and information in the form of tweets to improve the existing 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 perceptrons, a machine learning algorithm to analyse data. CNNs apply to image processing, natural language processing and other cognitive tasks. 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.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 12, December 2018

2. RELATED WORK

Traditional studies of cyberbullying were largely on a macroscopic view. These were conducted by social psychologists and scientists. However, these studies were largely focusing on the statistics of cyberbullying and concentrated more on the psychological way to prevent it. One introductory work has been presented in which several NLP models such as BOW, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) are applied to detect bullying signals in social media. Their result has verified the possibility of automatic cyberbullying detection.

A recent work of researchers in Massachusetts Institute of Technology proposed a novel methodology which is used to identify the bullying that happens over the domain of YouTube, i.e., the bullying which takes place in the form of YouTube comments are identified. Moreover, the proposed system is used to classify the YouTube comment into any of the following categories: sexuality, culture, intelligence and physical attribute. However, the outcome of the model was proven to be very less and provided ambiguity at times.

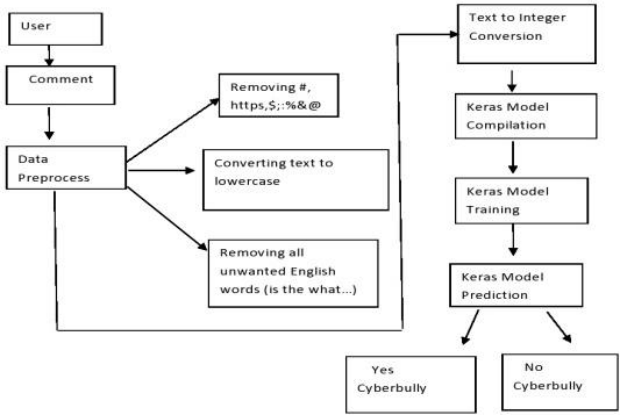
In other research work by Norton Online 2010 Malaysia for Detection of cyberbullying states that Malaysian children spend an average of 19 hrs a week on internet while that same also found that more than 80% of the children have been affected by negative comments and this is mainly done by their close friends or spouses. This work mainly focuses on the statistics of the cyberbullying.

In a research work Detection of cyberbullying is done by using naïve bayes technique. This method works fine for less data but for large data the outcome decreases. And also, this method gives best result in binary classification. One of the major problem with the existing methodology is that all the activities are focused more on the aftermath of the cyberbullying incident rather than an effective system to prevent the cyberbullying activities. Prevention is better than cure is the crux of our proposed architecture. Our paper targets on detecting the cyberbullying activities and classifying them into Cyberbully and non-Cyberbully which prevents the victims from facing the issues of cyberbullying and helps in taking preventive actions such as law enforcement, blocking or taking legal actions accordingly.

3. PROPOSED SYSTEM

In our proposed system, the notion of CNN implementation is included. 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. Inspired by the studies about the central nervous system of the mammals. A class of neural networks consist of significant number of layers of neurons, which are capable of learning by themselves is termed as deep learning. Deep learning in general consist of 3 layers,

•Input Layer
•Hidden Layer
• Output Layer



ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 12, December 2018

Our proposed model contains of a series of processes namely,
• Data Acquisition
• Data Pre-processing
• CNN Model Layers
• Model Prediction
• Obtaining Results

3.1 Data Acquisition

The foremost part of any project is getting the vital data. Thus, this is the crucial phase of extracting the proper dataset which is processed in the next stages into required information. In our proposed methodology, we tend to use twitter API and key token mechanisms to get the tweet id as well as the tweet message. This process is done around thousands of time to acquire considerable amount tweet data to train and test the dataset.

Example: Twitter_id- 552304521637285889
Tweet- Here is how #Islam inscribes the inferiority and abuse of women. (CYBERBULLYING)

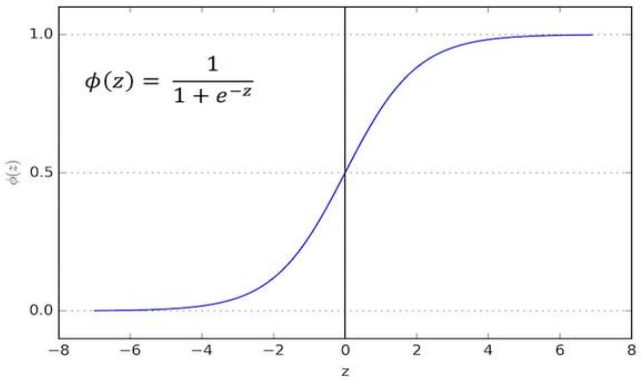
Twitter_id-552351551688564736
Tweet-Pete Evens looks orange #MKR. (NO CYBERBULLYING)

3.2 Data Pre-processing

Data pre-processing is cleaning of the data. It is the first and foremost step required in any process. It is the conversion of raw form of data into a required form of data for training the model in a proper manner. For example, in the raw form the data is “You look so ugly and fat #changethestyle”, after pre-processing the data is like “look ugly fat changestyle”. The pre-processed data takes out all the unwanted words like as, what, who, with, is, the etc. and special characters like @ () [] ?/; etc which are not required for training in the model. The data is separated into sentences and each sentence is made to make equal number of words by padding a common word which helps in uniformity of the data. Since the model accepts the data in the form of vector, the process makes the data into its lowercase format and converts that data into its vector form.

3.3 CNN Model Layers

The crux of the entire process depends upon the CNN layers used for processing. The main layers used in the model include Sequential Layer. The initial building block of keras is a model and the simplest model is called sequential model which consists of stack of neural network layers. The network is dense which means every node from each layer is connected with nodes from other layers. The perceptron is a single algorithm which takes the input vector x of m values as input and outputs either 1(yes) or 0(no) mathematically it is defined as f(x)=1 if wx+b>0 and f(x)=0 otherwise Perceptron is easy dealing with the small amount of data but in case of large data the perceptron is not helpful. That is, it cannot help in learning data. Since perceptron gives value either 0 or 1 the graph produced by it is discontinuous. We need something different and smoother. We need a function that progressively changes from 0 to 1 without any discontinuity.



ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 12, December 2018

Activation function can be of many types like sigmoid, ReLu etc. Sigmoid function is defined as 1/ (1+e− x) and can be used to produce continues values. A neuron can use the sigmoid for computing the nonlinear function z=wx+b where w is the weight of the neuron and b is the biased value. Activation function ReLu known as Rectified linear unit is also one such activation function which gives smooth values with nonlinear functions. A ReLu is simply defined as f(x)=max (0, x). The function is zero for negative values and grows gradually for positive values.

In our network we have converted the input text to a sequence of word indices. For that we have NLTK (Natural Language Toolkit) to parse the text into sentences and sentences to words. We could have used regular expression but statistical models for nltk are more powerful than regular expressions.

After creating the sequential model, the word indices are fed into array of embedding layers of a set size (in our case the longest word sentence) The output of the embedding layer is connected to the 1D Convolutional layer. This is then pooled into a single pooled word by a global max pooling layer. This vector is then input to a dense layer which outputs a vector (2) (Yes cyberbully and No Cyberbully). A SoftMax activation will return a pair of probabilies. The following shows our network model:

Embedding->Convolution1D->GlobalMaxpooling1D->Dense

3.4 Model Prediction

Once we define our model, we must compile it so that it can be executed by keras backend. (either Theano or Tensorflow). The model. compile consist of OPTIMIZERS, LOSS FUNCTION, METRICS. Optimizers are used to update weights while we train our model.

The process of optimization is defined as loss minimization. The loss functions can be of many type like MSE (Mean Squared Error), Binary Cross Entropy, Categorical Cross Entropy. In our case since we have binary classification we use Binary Cross Entropy. Suppose the model predicts p while the target is t, then the binary cross entropy is defined as: tlog(p)-(1-t)log(1-p)

On the other hand, categorical cross entropy is multiclass logarithmic loss. The object function is suitable for multiclass labels. The program goes like this: model. Compile (loss='categoricalcrossentropy 0, optimizer =0 adam0, metrics = ['accuracy'])

Some common choices of metrics are as follows:

- Accuracy: This is proportion of correct prediction with respect to targets.
- Precision: This is used to denote how many selected items are relevant for multilabel classification.
- Recall: Also used for multilabel classification.

Once our model is compiled it can be trained with the fit() function. The parameters used are:

- epochs- This is number of time model is exposed to training set.
- batchsize- This is number of training instances before optimizer performs a weight update.
- validationdata- This is the data that needs to be tested.

The following code explains the fit () method history=model.fit (Xtrain, Ytrain, batchsize = 64, epochs = 5, validationdata = (Xtest, Ytest))

3.5 Obtaining Results

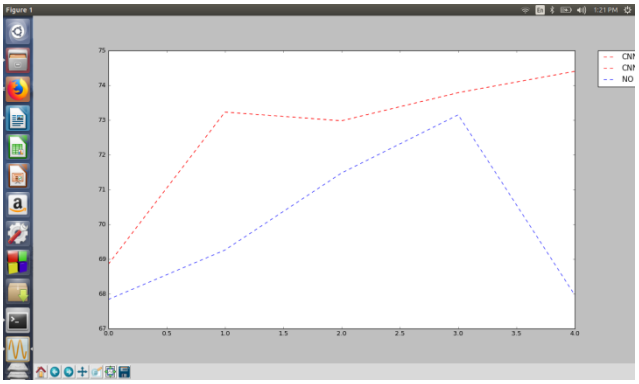
Moreover, we have depicted a visual representation of the results obtained with training the same dataset with two different algorithms. A comparison of the accuracy with NN versus CNN is plotted to provide a clear definition of why the existing system is less proficient than the proposed system. The proposed system also provides a novel idea of CNN implementation for text analysis. The system does not end with an analysis part, but our model also provides a prediction algorithm wherein the user input (in the form of tweet) is analysed with the existing training model and the output is provided in terms of percent of cyberbullying contained in the tweet. On providing a training model of over 12000 tweets, our model is tested to provide an accuracy of 75% approx., which proves the need for the proposed system.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 12, December 2018



4. RESULTS AND FUTURE ENHANCEMENTS

4.1 Results

Introduction of Twitter APIs which can be installed as web browser plugins to ensure that the cyberbullying tweets are prevented even before the user is notified about it. A fine grain approach of classification of cyberbullying over various categories can be effectively useful in a structured and efficient analysis of cyberbullying. These can include the classification of bullying under various categories like sexualism, racism, flaming. A mobile application to monitor and prevent the act of cyberbullying can be developed since most of the accessibility of social medium is through mobile phones. The domain of cyberbullying can be extended to other social media forums which includes Facebook.

Our proposed methodology has been proven to produce an accuracy of 75 percent and does provide consistent results over thousands of tweets analysed. However, when the same training set is used for a non-CNN analysis, the accuracy drops to 69 percent given the same amount of training and testing. According to our calculations for 1000 tweets we find TP, TN,FP,FN (which is True positive , True Negative, False Positive, False Negative and defined below). TP=850 TN=500 FP=250 FN=200

Table 1

Value	Accuracy	Prediction
True Positive	Yes	Yes
True Negative	No	No
False Positive	Yes	No
False Negative	No	Yes

Accuracy=((TP+TN)/TP+TN+FP+FN) *100

Recall=((TP/(TP+FN)) *100

Precision=(TP/(TP+FP)) *100

F=2*(Recall*Precision) / (Recall + Precision)

Table 2

Tweets	Accuracy	Recall	Precision	F
1000	75	80	77.27	78.61

DOI 10.17148/IJARCCE.2018.71215

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 12, December 2018

4.2 Future Enhancements

Since the domain of virtual bullying is a never-ending process, it is required that the methodologies require constant updating and upgrading to the current scenario. Our proposed methodology, can come useful in handling crisis situations and can even be enhanced to provide full-time support. Finally, it can even prevent a potential crisis. Some of the salient enhancements which can be inculcated soon include

5. CONCLUSION

In this paper, we propose a novel idea where any cyberbullying tweet remarks are identified as cyberbullying comment or not. The system uses a precise method of CNN implementation using keras and help in achieving accurate results. The proposed system can be used by the government or any organization- parents, guardians, institutions, policy makers and enforcement bodies. This can help the users by preventing them for becoming victims to this harsh consequence of cyberbullying.

REFERENCES

1. “Text classification using convolution neural networks. (2017).
2. Keras tutorial deep-learning in python.
3. B. Sri Nandhinia, J. (2015). “Online social network bullying detection using intelligence techniques.
4. K. Dinakar, R. R. and Lieberman, H. (2011). “Modelling the detection of textual cyberbullying.
5. K. Reynolds, A. K. and Edwards, L. (2011). “Using machine learning to detect cyberbullying.
6. Mohammed Ali Al-garadi*, Kasturi Dewi Varathan, S. D. R. (2016).
7. “Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network.
8. Rui Zhao, Anna Zhao, K. M. “Automatic detection of cyberbullying on social networks based on bullying features.
9. V. Nahar, X. L. and Pang, C. (2013). “An effective approach for cyberbullying detection.
10. Whittaker, E.,. K. R. M. (2015). “Cyberbullying via social media.
11. Archer (2018) B. Sri Nandhinia (2015) Mohammed Ali Al garadi* (2016)
12. Rui Zhao (Rui Zhao) K. Dinakar and Lieberman (2011) K. Reynolds and Edwards (2011) Whittaker (2015) V. Nahar and Pang (2013) lin (2017)
13. Tweet classification of sentimental analysis using keras in python
14. Deep Learning for detecting cyberbullying across social media platforms by S Agarwal A Awekar.
15. Detecting state of aggression in sentence By R potapova
16. Hate speech detection on Facebook (Blog)
17. Analytics Vidya (Website for python and CNN)