Detection of Cyberbullying Using Deep Neural Network

Vijay Banerjee
Computer Engineering
Viva Institute of Technology
Virar, India
banerjeevijay1996@gmail.com

Jui Telavane
Computer Engineering
Viva Institute of Technology
Virar, India
jui.telavane17@gmail.com

Pooja Gaikwad
Computer Engineering
Viva Institute of Technology
Virar, India
gaikwadpoo67@gmail.com

Pallavi Vartak
Computer Engineering
Viva Institute of Technology
Virar, India
Pallavivartak@viva-technology.org

Abstract— **Innovation is developing quickly today. This headways in innovation has changed how individuals cooperate in an expansive way giving communication another dimension. But despite the fact that innovation encourages us in numerous parts of life, it accompanies different effects that influence people in a few or the other way. Cyberbullying is one of such effects. Cyberbullying is a wrongdoing in which a culprit focuses on an individual with online provocation and loathe which has antagonistic emotional, social and physical effects on the victim. So as to address such issue we proposed a novel cyberbullying detection method dependent on deep neural network. Convolution Neural Network is utilized for the better outcomes when contrasted with the current systems.**

Keywords— **cyberbullying; detection; deep learning; convolutional neural network; word embedding, Glove.**

I. INTRODUCTION

As the Technology advances people rely on the technolgy to conductmany of the daily activities including communication. The social media platforms are thus a great platform connecting people all over the world. But this promising shift towards digital world comes with a pricy cost. The secrecy over the web has made it less demanding for individuals to post audit remarks on various issues without being identified, a precedent is found in the informal organization twitter. In this manner the client can post tweets which can assault different clients, and these tweets are untraceable. This training prompts cyberbullying. Cyberbullying is characterized as any fierce, purposeful activity directed by people or gatherings, utilizing on the web channels over and again against a victim who does not can possibly react[1]. Numerous investigations have tended to cyberbullying with point of surveying its commonness, and results demonstrated that cyberbullying is a typical issue confronting the present age and that the quantity of victims is rising[2][3].

So as to help in controlling cyberbullying and restricting its commonness, numerous cyberbullying recognition instruments have been presented. There are numerous exisiting framework on cyberbullying location and the examination is significantly more developed yet numerous issues are not tended to yet. Most of research on cyberbullying location centers around spaces, for example, machine learning and data mining. To the extent the investigation none of the earlier research has tended to the issue of digital assault identified with network, Sexism and Racism utilizing profound neural network. In the proposed system, we use convolution neural network to create a model that detect bully related tweets and predict the behaviour of the new data introduced. The proposed approach use word vectors that are feed to the CNN for classification of tweets. The dataset of chats and tweets form various social media platforms is collected for the evaluation.

II. RELATED WORKS

An detail literature survey conducted across IEEE explorer and Springer digital library. The main search strategy was the discovery of academic literature relavent to the theme “Techniques of cyberbullying detection”. The following is the detailed literature work.

Rui Zhao and Kezhi Mao [5] used a new representation learning method. This method is Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA) developed via semantic extension of the popular deep learning model stacked denoising autoencoder. Elaheh Raisi and Bert Huang [6] proposed 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. [7] 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. A. Conneau, et. al. [8] presents a new architecture (VD-CNN) for text processing which works specifically at the character level and uses as it

978-1-5386-9533-3/19/$31.00 ©2019 IEEE

were small convolutions and pooling operations.

S. Bhoir, et. al. [9] presented a similar investigation of various word embedding techniques based on various parameters. classifier used and effect of changes in dimensionality. E. Raisi, et. al. [10] presented the participant vocabulary consistent model a weakly supervised methodology for concomitantly learning the roles of online social users in the badgering type of cyberbullying. H. Zeng, et. al. [11] utilized a representation procedure that has 4 connected view that helps to investigate learning parameters. It is important to get the knowledge of how the model parameters advance from lower to higher precision so we can improve the training procedure. V. N. Kumar, et. al. [12] This paper uses 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

Andrew M. Dal and Quoc V. Le [13] proposed sequence learning supervised model using CNN and LSTM.K. Duan, et. al. [14] explains SoftMax combination for multicategory classification both one-versus-all and one-versus-one classifier. This paper explains how to efficiently extend binary classification method for multi-category classification. Q. Li, et. al. [15] proposed a new tweet sentiment classification approach using SSWE and WTFM produce classes based on the weighting scheme and text negation and a new text classification method. I. Raid [16] carried out research using data mining technique here there are several stages as data collection, preprocessing, TF-IDF, weighting, data validation and classification using naive Bayes classifier. K. Sahay, et. al. [17] explains the training in machine learning model using supervised learning.

III. Proposed Methodology

CNN is a Type of Deep Neural Network Model that comprises of neurons. Neurons work on biases and learnable weights. The architecture of CNN comprises of an input, output layer and a few of hidden layers.The input layer includes succession of vectors. It is examined utilizing fixed size of filter. The filter shifts or strides only one row or one column on the matrix. Each filter distinguishes different features in the content so as to portray it into the feature map the next layer is maxpooling layer. The maxpooling layer minimizes the features in the feature map. It selects the highest value in the feature map to capture most important feature. Due to that, it decrease the calculation in the propelled layers the dropout strategy is connected to diminish overfitting with dropout rate is 0 5. The last layer of the model is Dense(FullyConnected) layer and it is used for classification purpose. This layer classifies the text based on the classes specified.

A. System Flow

For a proposed solution the flow diagram is as follows:

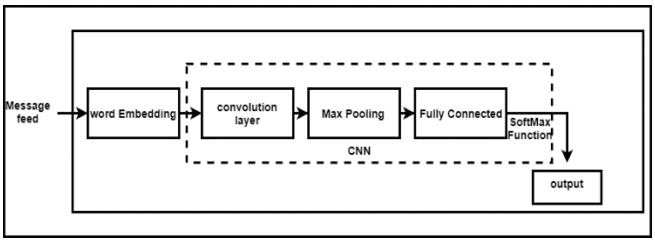

fig 1. 1 : System flow

B. Word Embedding

Starting word embeddings choose information portrayal for Neural Network models. In any case, amid the training, Neural Network model adjust these underlying word embeddings to learn assignment specific word embeddings. For our proposed system we experimented with different word embedding methods and used Glove for our project.

C. Architecture

The very next layer from the embedding layer (if there should arise an occurrence) is the convolutional layer. This is the main core layer of the whole convolution neural network. In our project we have used different layers of convolution as follows:

- Input
- Word Embedding layer
- Dropout layer(0.25)
- Convolution layer
- Max Pooling
- Convolution-layer
- Max Pooling
- Convolution layer
- Max Pooling
- Dropout layer(0.5)
- Fully connected layer
- Softmax
- Classification layer

Input layer:
in this layer data from the tweeter dataset is given to the embedding model.

Embedding layer:

this word embedding is done for words into vector conversion. Neural Network only takes numerical value as input so, tokens that is words are converted into vectors thus it each represents text as a row of vectors Each word in the text, which is one token is embedded into a vector. This step is a

matrix of size m x n, where n is the length of the vector and m is the number of tokens in the texts.

Dropout layer:

Dropout is a regularization technique that approximates training an expansive number of neural networks with various structures in parallel. In proposed arrangement dropout layers are utilized subsequent to embedding and before fully connected layer for regularization.

Convolutional Layer

Each input contains an arrangement of vectors. It is checked utilizing fixed size of filter. The filter moves or strides just a single row or one column over the matrix. Each filter identifies various features in the content so as to speak to in the element map.

Rectified Linear Unit (ReLU)

Layer applies a max function f(x)=max(x,0) to the matrix of the convolved after convolution. It sets all the negative qualities in the dot products of the matrix to 0. Every other esteem are unaltered. It builds the speed of training the network by evacuating negative activation in the slope, consequently staying away from complex negative computations.

Max-Pooling Layer

After convolutional layer, the following layer is Max-Pooling layer. The Max-Pooling layer limits the features in the Feature map. It chooses the most noteworthy incentive feature in the feature map to catch most critical feature.

The proposed system Architecture is as follows:

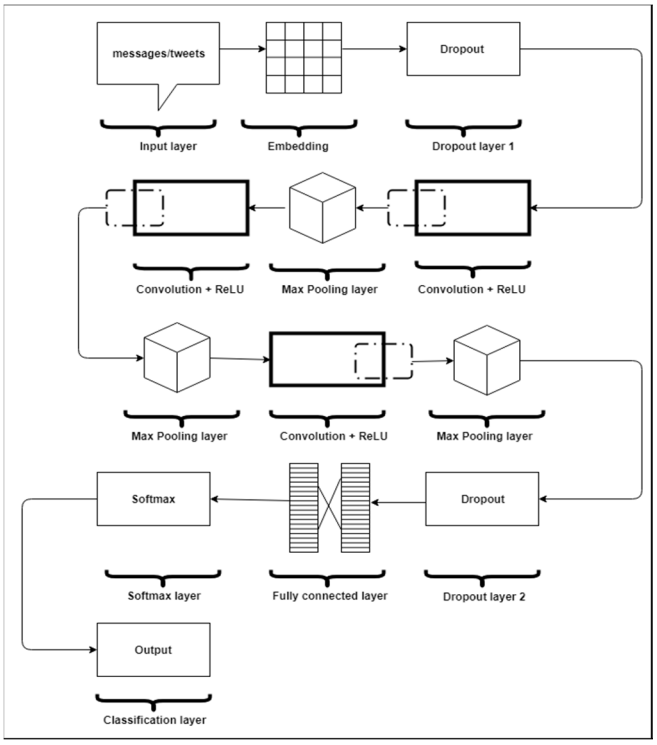

fig 1. 2 Model Architecture

Fully Connected Layer

The last layer of the model is Dense (FullyConnected) layer and it is utilized for classification reason. This layer groups the content dependent on the classes determined.

IV. 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 deeplearning. The Twitter dataset used consists of 69874 tweets, which are converted to vectors using open source wordembedding Glove. These messages were sorted and labels were generated.

Neural Network model revealed here were implemented utilizing Keras on top of tensorflow. We preprocess 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.

IV. Result and Analysis

The below Table 1, shows the results of the various existing cyberbullying detection systems based on data mining, machine learning and RNN based deep learning system. The table shows that the proposed system derives better results than all the systems referred so far.

Title of paper	Techniques used	Accuracy
Optimized Twitter Cyberbullying Detection based on Deep Learning [4]	RNN, GloVe	81.60%
Detecting cyberbullying and aggression in social commentary [17]	Natural Language Processing and Machine Learning	75%
Detection of cyberbullying on social media using data mining techniques [16]	Data Mining	75%
Detection of Cyberbullying using deep neural network(Proposed System)	**CNN (Convolution Neural Network)**	**Testing Accuracy-93.97%**

Table 1. Result Analysis

VI. Conclusion And Future Scope

Rapid growth of technology is affecting the way we communicate on the social media platforms resulting in

cyberbullying and many such issues. although many researches addressed cyberbullying in SMP(social medai platform), The techniques used for the detection proves inefficient in classification. In the proposed 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 regressive training of system so as to detect cyberbullying in real time chats and also the detection of cyberbullying in chats containing Hinglish(hindi and english) code mix languag.

REFERENCES

1. R. Shetgiri, "Bullying and Victimization Among Children", Advances in Pediatrics, vol.60, no. 1, pp. 33-51, 2013.
2. Brown, E. Clery and C. Ferguson, "Estimating the prevalence of young people absent from school due to bullying," National Center for Social Research, 2011.
3. Van Royen, K. Poels, W. Daelemans and H. Vandebosch, "Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability" Telematics and Informatics, vol. 32, no.1, pp.8997, 2015.
4. Monirah A. Al-Ajlan, Mourad Ykhlef, “Optimized Twitter Cyberbullying Detection based on Deep Learning”, 978-1-5386-4110-1, IEEE-2018.
5. Rui Zhao,Kezhi Mao “CyberBullying Detection based on Semantic-Enhance Marginalize Denoising Auto-encoders”IEEE Transaction on Affective Computing, 2015.
6. Elaheh Raisi, Bert Huang “Weakly Supervised Cyberbullying Detection with Participant Vocabulary Consistency” Social Network Analysis and Mining, May 24,2018.
7. 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.
8. Alexis Conneau, Holger Schwenk, Yann Le cun “Very Deep CNN for Text Classification”Association for Computational Linguistics, Volume1, pages 1107-1116,7 April 2017.
9. MS.Snehal Bhoir, Tushar Ghorpade, Vanita Mane “Comparative Analysis of Different Word Embedding Models” IEEE,2017.
10. Elaheh Raisis,Bert Huang “Cyberbullying Detection with Weakly Supervised Machine Learning” International Conference on Advances in Social Networks Analysis andMining IEEE/ACM,2017.
11. Haipeng Zeng, Hammad Haleem, Xavier Plantaz, NanCao and Huamin Qu “CNN Comparator: Comparative Analytics of CNN” arXiv,15 Oct,2017.
12. Vandana NandaKumar,Binsu C,Kovoor,Sreeja M.U “Cyber-Bullying Revelation in Twitter Data using Naive-Bayes Classifier Algorithm” International Journal of Advanced Research in Computer Science. Volume 9, No. Jan-Feb 2018.
13. Andrew M.Dal,Quoc V.Le “Semi-Supervised Sequence Learning “ arXiv,4 Nov 2015.
14. Kaiob Duan, S.Sathiya Keerthi,Wei Chu, Shirish Krishnaj Shevade and Anu Neow Poo“Multi-Category Classification by Softmax Combination of Binary Classifiers” Department of Computer Science and Automation, Bangalore.
15. Quanzhi Li, Sameena Shah, Rui Fang, Armineh Nourbakhsh, Xiaomo Liu “Tweet Sentiment Analysis by Incorporating Sentiment Specific Word Embedding and Weighted Text Features” International Conference on Web Intelligence IEEE/WIC/ACM,2016.
16. Hariani,Imam Raid “Detection of Cyberbullying on Social Media using Data Mining Techniques” International Journal of Computer Science and Information Security, Vol.15, No.3, March 2017.
17. Kahitiz Sahay, Harsimran Singh Khaira,Prince Kukreja, Nishchay Shukla “Detecting Cyberbullying and Aggression in Social Commentary using NLP and Machine Learning” International Journal of Engineering Technology Science and Research, ISSN-2394 3386, Volume5, Issue1, January 2018