IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CYBER BULLYING DETECTION

Roshni Jadhav, Grisha Chaudhari, Sumeet Rane

Information Technology Department,
Atharva College of Engineering
Mumbai, India

Abstract: The advancement of technology, craze of social networking platforms is proliferating. Online users now share their information with each other easily using computers, mobile phones etc. However, this has lead to the growth of cyber-criminal acts for example, cyberbullying which has become a worldwide epidemic. It has emerged out as a platform for insulting, humiliating a person which can affect the person either physically or emotionally and sometimes leading to suicidal attempts in the worst case. To overcome this problem, many methods and techniques had been worked upon till now to control this problem. Cyberbullying is a disturbing online misbehavior with troubling consequences. It appears in different forms, and in most of the social networks, it is in textual format. In recent studies, deep learning based models have found their way in the detection of cyberbullying incidents, claiming that they can overcome the limitations of the conventional models, and improve the detection performance. Existing works for cyberbullying detection have at least one of the following three bottlenecks. First, they target only one particular social media platform (SMP). Second, they address just one topic of cyberbullying. Third, they rely on carefully handcrafted features of the data. We show that deep learning based models can overcome all three bottlenecks. Knowledge learned by these models on one dataset can be transferred to other datasets. We performed extensive experiments using real-world datasets: Twitter.

Index Terms **- Cyberbullying Detection, Text Mining, Representation Learning, Word Embedding.**

I. INTRODUCTION

Cyberbullying is the use of electronic communication to bully a person by sending harmful messages using social media, instant messaging or through digital messages. It has emerged out as a platform for insulting, humiliating a person which can affect the person either physically or emotionally and sometimes leading to suicidal attempts in the worst case. The main issue in preventing cyberbullying is detecting its occurrence so that an appropriate action can be taken at initial stages. To overcome this problem, many methods and techniques had been worked upon till now to control this problem. With the emergence of Web 2.0 there has been a substantial impact on social communication, and relationships and friendships have been redefined all over again. Adolescents spend a considerable amount of time online and on different social platforms, besides all the benefits that it might bring them, their online presence also make them vulnerable to threats and social misbehaviors such as cyberbullying. Cyberbullying needs to be understood and addressed from different perspectives. Automatic detection and prevention of these incidents can substantially help to tackle this problem. There are already tools developed which can flag a bullying incident [4] and programs which try to provide support to the victims [5]. Moreover, most of the online platforms which are commonly used by teenagers have safely centers, for example, YouTube Safety Centre4 and Twitter Safety and Security5, which provide support to users and monitor the communications. There are also many research conducted on automatic detection and prevention of cyberbullying, which we will address in more details in the next section, but this problem is still far from resolved and there is the need for further improvements towards having a concrete solution. Most of the existing studies [6]–[9] have used conventional Machine Learning (ML) models to detect cyberbullying incidents. Recently Deep Neural Network Based (DNN) models have also been applied for detection of cyberbullying [10], [11]. In [11], authors have used DNN models for detection of cyberbullying and have expanded their models across multiple social media platforms. Based on their reported results, their models outperform traditional ML models, and most importantly authors have stated that they have applied transfer learning which means their developed models for detection 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 which is adaptable and transferable to different social networks is of great value. We expand our work by re-implementing the models on a new dataset. For this purpose, we have used a Twitter dataset which has been extensively used in cyberbullying studies [6], [15], [16]. The ultimate aim was to investigate the interoperability and the performance of the reproduced models on new datasets, to see how adaptable they are to different social media platforms and to what extent models trained on a dataset (i.e., social network) can be transferred to another one. This provides a base to compare the outcome of DNN models with the conventional ML models.

II. RELATED WORK

Cyberbullying is recognized as a serious incident at least since 2003 [13]. The use of social media shatters violently 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 situation as offline bullying [10]. However, the automatic detection of cyberbullying was addressed only since 2009 [19]. According to the research topic, cyberbullying detection is a text classification problem. Existing works fit in the following template: get training dataset from single Social Media Platform (SMP), engineer a variety of features with a certain style of cyberbullying as the target, apply traditional machine learning methods, and evaluate the success of cyberbullying in terms of measures such as F1 score and accuracy. These works heavily rely on handcrafted features of models such as the use of swear words. These methods tend to have low precision for cyberbullying detection as handcrafted features of models are not strong against dissimilarity in bullying style across (SMPs) and bullying topics. Recently, deep learning methods have been applied for cyberbullying detection [2].

Rui Zhao and Kezhi Mao [1] used a new representation learning method that has been proposed to tackle this problem. This method is Semantic-Enhanced Marginalized Denoising Auto Encoder (smSDA) developed the popular deep learning model stacked denoising auto encoder. This method is able to exploit the feature of bullying input information and learn a robust and discriminative representation of bullying text.ElahehRaisi and Bert Huang [2] proposed a weakly supervised machine learning method for simultaneously playing user roles in harassment-based bullying and vocabulary of bullying. The learning algorithm considers a social structure and infers which users tend to bully and which tend to be victimized. P. Zhou, et. al. [3] Proposed Attention-Based Bidirectional Long Short-Term Memory Networks (Att- BLSTM) to capture the most important semantic information in a sentence. This paper is using BLSTM with an attention mechanism for text classification, which can be automatically priority on the words that have a strong effect on classification, to capture the most important semantic information in a sentence of bullying comment, without using extra knowledge and NLP systems. E. Raisi, et. al. [7] presented the participant-vocabulary consistent model a weakly-supervised approach for simultaneously learning the roles of social media users in the harassment form of cyberbullying and the tendency of language indicators to be used in such cyberbullying. K. Sahay, et. al. [16] 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 the regular user the NLP and machine learning are studied and evaluated for the task of identifying bullying comment in the dataset. As the use of the internet and social media increases the issue of cyberbullying is becoming subjugated. To solve this problem cyberbullying detection is necessary. The existing system uses machine learning and data mining techniques for the detection of cyberbullying. But these approaches limit the correctness of the detection and work only on some limited classified features of cyberbullying. The previous work also often uses word embedding techniques for data representation. To overcome this problem, we use neural networks and deep learning which gives better results. The advantage of this is that using neural networks helps to work with a large amount of data easier. The system uses CNN neural network models for efficient and perfect results.

III. DEEP NEURAL NETWORK BASED MODELS

DNN models are used for the detection of cyberbullying: Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BLSTM). These models are respectively different in complexity in their neural architecture. CNN's are mostly used for image and text classification [7], [8] as well as sentiment classification [9]. LSTM networks are used for learning long-term dependencies. Their internal memory of models makes these networks useful for text classification [12], [2]. Bidirectional LSTMs, increase the input information of the data to the network by encoding information in both forward and backward direction [2]. BLSTM with attention gives a more direct relationship between the state of the model at different points in time [3]. Most of the models have identical layers except for the neural architecture layer in which is unique to each model. The embedding layer, which will be explained in more detail in following, processes a fixed-length sequence of words. There are two dropout layers that are used to avoid overfitting, once before (with dropout rates of 0.25) and one after (with dropout rates of 0.5) the neural architecture layers. Then there is the fully connected layer which is a dense output layer with the number of neurons equal to the number of classes.

3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are known to have a good performance on data with a high locality when words get more care weight about the features surrounding them. We are trying to get high priority in the text given their short length and their tendency to focus on cyberbullying. We used CNNs that received input text in the form of sequences of integer representations are arising from the unigrams. The character processing included the conversion of emoticons into word 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 using a dictionary of integers, with the original ordering of the tokens preserved.[5][8].

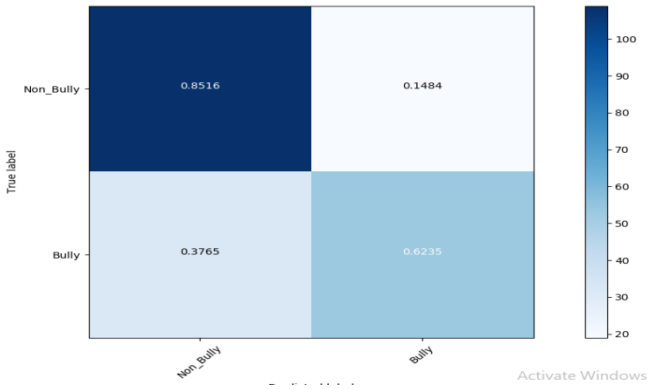

Fig 1. Confusion Matrix of CNN

The confusion matrix shows the ways in which your classification model is confused when it makes predictions.
Predicted Label
True Positive = 0.8516 False Negative = 0.1484
False Positive = 0.3765 True Negative = 0.6235

3.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a type of periodic neural network capable of learning order dependence in sequence prediction problems. This is a behaviour required domains like machine translation, speech recognition, and more. It can be difficult to get your hands around what LSTMs are, and how terms like bidirectional and sequence-to-sequence relate to the field. LSTM Network used to sort and making predictions on words based on time series data since there can be lags of duration between important events in a time series [3]. LSTMs are developed to deal with explode and vanish gradient problems that can be encountered when training traditional RNNs. Relative insensitivity is an advantage of LSTM over RNNs, hidden Markov models and other sequence learning methods and models in numerous applications.

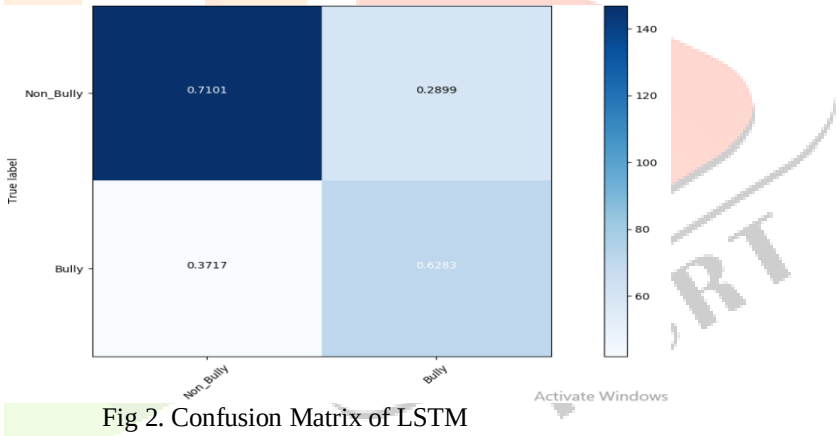

Fig 2. Confusion Matrix of LSTM

The confusion matrix shows the ways in which your classification model is confused when it makes predictions.
Predicted Label
True Positive = 0.7101 False Negative = 0.2899
False Positive = 0.3717 True Negative = 0.6283

3.3 Bidirectional LSTM

Bidirectional LSTMs are an extension of LSTMs that can improve model performance on sequence classification problems. In problems where all time steps of the input sequence are available, Bidirectional LSTMs train two instead of one LSTMs on the input sequence. This can provide additional input context to the network and result in faster and even fuller learning on the problem. It involves duplicating the first periodic layer in the network so that there is now two layers side-by-side, then providing the input sequence as-is as input to the first layer and providing a reversed copy of the input sequence to the second layer. The use of sequence bi-directionally was initially justified in the domain of speech recognition because there is evidence that the input context of the whole utterance is used to interpret what is being said rather than a simple interpretation. The use of bidirectional LSTMs may not make sense for all prediction problems but can offer benefits in terms of better results to those domains where it is appropriate.

IV. PROPOSED SYSTEM

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.

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

Convolution Neural Network

Vectors generated by word embedding are 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.

System Flow

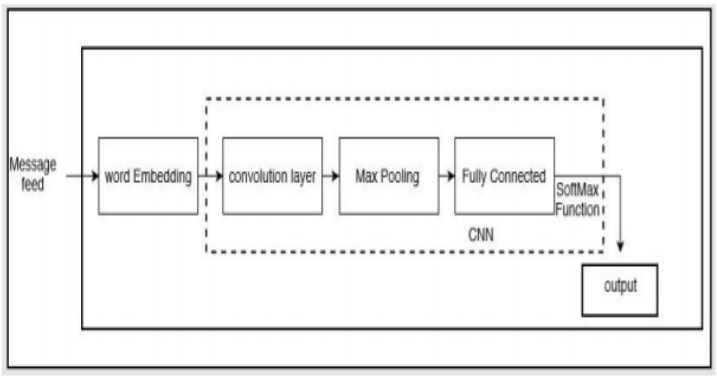

Fig.3. Block Diagram of Word Detection

Long short Term Memory networks, LSTMs have been observed as the most effective solution. LSTMs have an edge over conventional feed-forward neural networks and RNN in many ways. This is because of their property of selectively remembering patterns for long durations of time. LSTMs, make small modifications to the information by multiplications and additions. With LSTMs, the information flows through a mechanism known as cell states. This way, LSTMs can selectively remember or forget things. The information at a particular cell state has three different dependencies. The information at a particular cell state has three different dependencies.

These dependencies can be generalized to any problem as:

1. The previous cell state (i.e. the information that was present in the memory after the previous time step)
2. The previous hidden state (i.e. this is the same as the output of the previous cell)
3. The input at the current time step (i.e. the new information that is being fed in at that moment)

Another important feature of LSTM is its analogy with conveyor belts!
Industries use them to move products around for different processes. LSTMs use this mechanism to move information around. We may have some addition, modification or removal of information as it flows through the different layers, just like a product may be molded, painted or packed while it is on a conveyor belt.

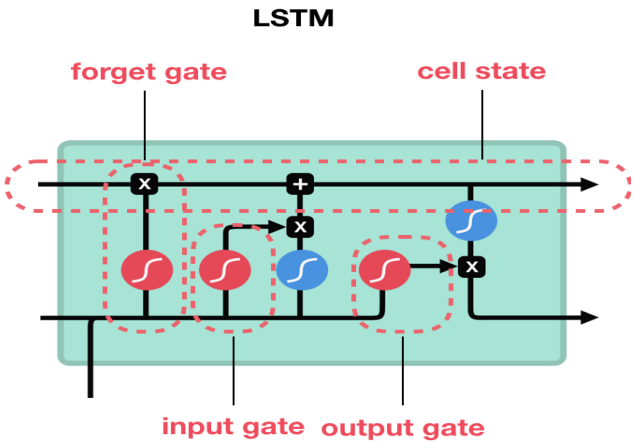

Fig 4. LSTM cells & Its Operations

A typical LSTM network is consists of different memory blocks called cells There are two states that are being data transferred to the next cell; the cell state and the hidden state. The memory blocks are responsible for retrieval things and manipulations to this memory are done through three major mechanisms, called gates. Forget Gates, Input Gates, Output Gates.

Forget gate
This gate decides which information should be kept.

Input Gate
To update the cell state, we have the input gate. That decides which values will be updated by transforming the values between 0 and 1. 0 means not important and 1 means important.

Cell State
First, the cell state gets pointwise multiplied by the forget vector. This has the possibility of dropping values in the cell state if it gets multiplied by values near 0. Then we take the output from the input gate and do a pointwise addition which updates the cell state to new values that the neural network finds relevant. That gives us our new cell state.

Output Gate

The output gate decides what the next hidden state should be in the cell. The hidden state is also used for predictions. The output is the hidden state.

V. RESULT & ANALYSIS

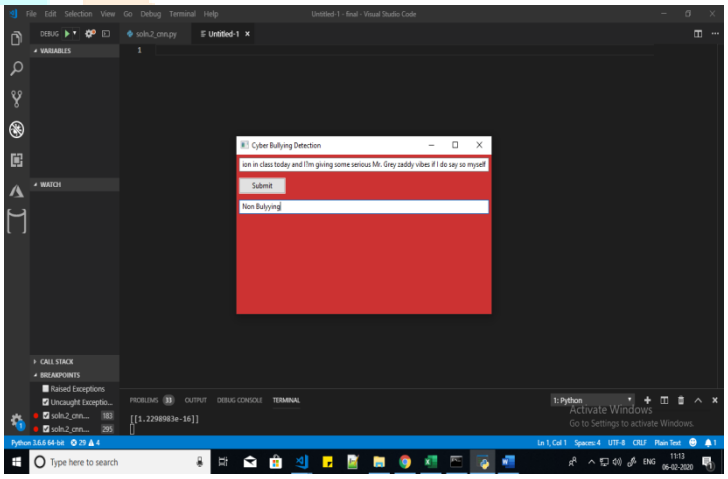

Fig 5. Non-Bullying Comment Detected

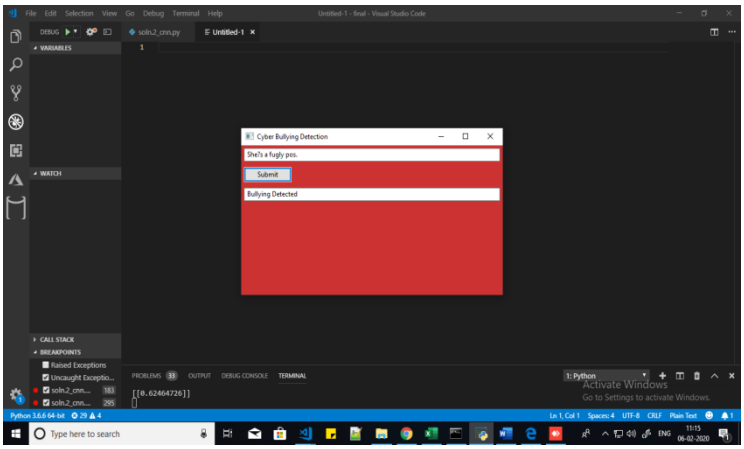

Fig 6. Bullying Comment Detected

TABLE I: RESULT ANALYSIS

Techniques Used	Accuracy
RNN, GloVe	60%
NLP	57%
Data Mining	57%
CNN, LSTM	66%

VI. CONCLUSION

The proposed system will serve as an ideal model for the right detection of cyberbullying posts on the social media platform's thus overcoming various downfalls in the process of detection existed till days. Further proposed system also makes use of efficient training models and word embedding methods that makes the system novel. The system proves to be useful for the analysis of the cyberbullying rates on different social media platforms so that relative precautions and actions can be taken to decrease the cyberbullying rate..

References

[1] Rui Zhao,Kezhi Mao "CyberBullying Detection based on Semantic - Enhanced Marginalized Denoising Auto-encoders" IEEE Transaction on Affective Computing, 2015.

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

[3] 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.

[4] 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

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

[6]] MS.Snehal Bhoir,Tushar Ghorpade,Vanita Mane "Comparative Analysis of Different Word Embedding Models" IEEE,2017.

[7] Elaheh Raisis,Bert Huang "Cyberbullying Detection with Weakly Supervised Machine Learning" International Conference on Advances in Social Networks Analysis and Mining IEEE/ACM,2017.

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

[9] 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.

[10] Andrew M. Dal and Quoc V. Le," Learning Longer-term Dependencies in RNNs with Auxiliary Losses" Proceedings of the 35th International Conference on Machine Learning,Stockholm, Sweden, 13 June, 2018.

[11] Duan K., Keerthi S.S., Chu W., Shevade S.K., Poo A.N. (2003) Multi-category Classification by Soft-Max Combination of Binary Classifiers. In: Windeatt T., Roli F. (eds) Multiple Classifier Systems. MCS 2003. Lecture Notes in Computer Science, vol 2709. Springer, Berlin, Heidelberg.

[12] Q. Li, 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.