Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

ISSN NO. 2320-5407

Article DOI: 10.21474/IJAR01/11266
DOI URL: http://dx.doi.org/10.21474/IJAR01/11266

RESEARCH ARTICLE

CYBERBULLYING DETECTION & PREVENTION ON SOCIAL NETWORKS

Diganto Deb Barma[1], Vishal Dwvedi[1] and Akshatha Ballal[2]

1. Final Year Student of B.E, Department of Information Science & Engineering, Acharya Institute of Technology, Bangalore-560090, Karnataka, India.
2. Assistant Professor, Department of Information Science & Engineering, Acharya Institute of Technology, Bangalore-560090, Karnataka, India.

……………………………………………………………………………………………………....

Manuscript Info

……………………….

Manuscript History
Received: 05 May 2020
Final Accepted: 10 June 2020
Published: July 2020

Key words:-
Cyberbullying, Cybercrime, Deep Learning, Twitter

Abstract

………………………………………………………………

As social media becomes, more popular cyberbullying has become a household word nowadays. Before the revolution of social media, cyberbullying is generally controlled by following standard guidelines, the use of human technicians, and blacklisting disrespectful words. Nevertheless, after the revolution, these mechanisms fall short for handling billions of users on social media. So, there is a need for building a system that can detect the sentiment of the content posted by the users and prevent the material from posting based on the sense of the content. However, it is difficult because the content from social media is mostly unstructured, short, and noisy, and frequently there's a use of confusing abusive phrases and words. To make cyberspace a secure place for future generations, a cyberbullying detection & prevention system is proposed.

……………………………………………………………………………………………………....

Introduction:-

Cyberbullying is an act of intentionally & repeatedly hurting others by using abusing & disrespecting words. Cyberbullying behavior is a common phenomenon among the young generation nowadays, which is a dangerous problem. Cyberbullying behavior is increasing rapidly due to easily accessible social media platforms. As there are limited guidelines in terms of maintaining this type of act, users can express anything about anyone on this platform. This type of behavior leads to various psychological disorders, i.e., depression, anxiety, trauma, and suicide.

There is a mandatory need for a system capable of identifying the sentiment of contents to prevent cyberbullying behaviors in the social world. There are several ways to deal with cyberbullying, including researching facts associated with it and educating people from the ground level. The traditional method for detecting cyberbullying is a human moderator. These moderators track the users who do not follow the guidelines and blacklisting them based on the guidelines. It cannot scale well when we have a broader audience. Recent developments have also built models where it automatically detects cyberbullying content, but no action is taken. The accuracy of the model is not up to the mark as the data is unstructured and also limited in our best knowledge. Existing industry tools allow us to filter or report harmful content only.

Corresponding Author:- Diganto Deb Barma
Address**:-** Final Year Student of B.E, Department of Information Science & Engineering, Acharya Institute of Technology, Bangalore-560090, Karnataka, India.

Literature Review:-

In they have demonstrated how machine learning algorithms can be used to identify abusive words. Naïve Bayes classifier and various Support Vector Classifiers, i.e., Gaussian, SVC with linear kernel, and LinearSVC, have been used-among them SVC performed the best. Syntactic patterns and semantic information, along with POS tagging, can be implemented for optimizing precision and recall.

In, the author's primary focus is to identify cyberbullying performers based on the users' texts and integrity analysis. To detect a text containing abusive words, they have followed eight standard guidelines for feature extraction. The system learned the pattern of the feature from KNN and SVM algorithms.

In Support Vector Machine (SVM) is used to classify the Facebook post to identify cyberbullying behaviors. The TF_IDF result was used to measure the importance of each post. Most frequent words were identified by Selenium that, in turn, used to output a new set of Facebook posts. The error matrix has been computed to measure the accuracy of the SVC model.

The semantic of the word is maintained by representing a tweet as a set of word vectors. Data is collected via twitter4j API, and java code was written to fetch random tweets. Word embedding was used to identify the similarities between the words, and CNN was used for classification of the tweets.

A socio-linguistic model that jointly detects cyberbullying content in messages identifies participant roles and exploits social interactions. A collective probabilistic approach is proposed in order to find dependencies between language and participants in social interactions. One limitation is that only one social media platform is investigated. It is the first model in this domain which jointly infers bullying content and participant roles.

A cyberbully detection method based on deep neural networks (Convolutional Neural Network) is proposed. The proposed system is implemented in Python and Tensorflow. Based on the detailed analysis, it has been shown that the proposed system outperforms all other traditional systems.

Detection of cyberbullying is done on Arabic Tweet messages based on the strength of the bullying message. Twitter API uses to access the real-time activities of a user via webhooks. The system could be used by parents to monitor their kid's activities. The system could be improved by detecting cyberbullying based on the context of the posts.

Proposed system:

Figure 1 provides a flowchart of a system that is built to detect the sentiment of the text & detected sentiment is used as a prevention mechanism. The model is deployed on a blogging website using Flask. A user first needs to register themselves or login if already registered. After authentication, the user moves to the homepage of the blogging site. The user can post content, update the content whenever needed, and delete the content is unnecessary. Whenever a user tries to post something, the content of that post is sent to the backend, where the model will classify the sentiment of the content. Based on the classification result, the content is posted on the blog site if it is found as non-offensive; otherwise, it is restricted from posting, and an error is shown to the user.

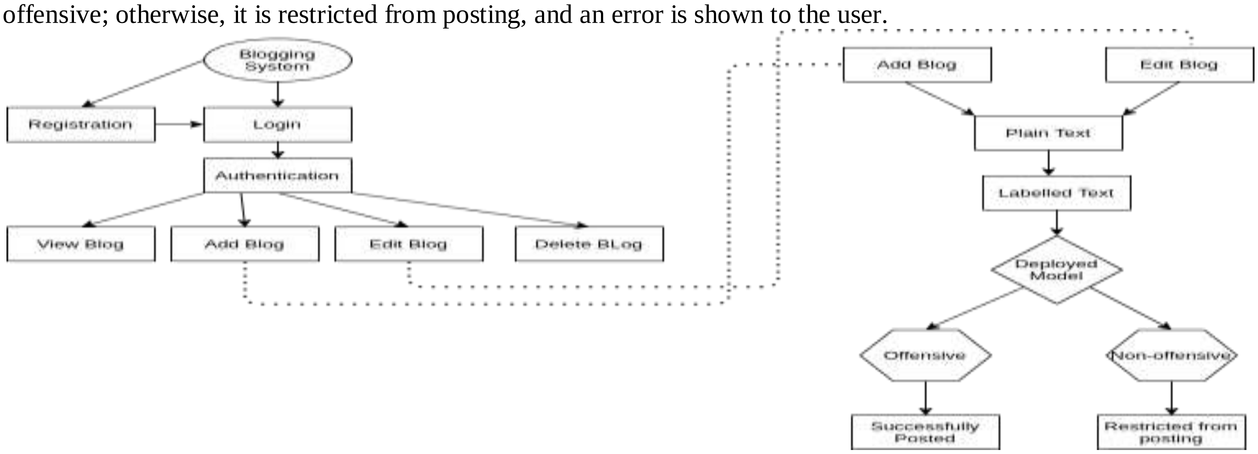

Figure 1:- Flowchart of the system.

The dataset is collected from twitter API. The collected data is in the text from which contains much unnecessary information such as hashtag, punctuation numbers. As machines only understand numbers, we need to remove this unnecessary information and convert it into numbers so that the meaning of the original text remains the same. To do this, we have followed the below steps:

1. Removing punctuation
2. Tokenization
3. Removing stopwords
4. Stemming

Removing punctuation:
Punctuation marks are used to accomplish the meaning of a sentence. Our concern is to understand the sentiment of a particular sentence, which can still be achievable without punctuation marks. Moreover, punctuation marks add noise to the dataset, impacting the learning process while training the model.

Tokenization:
Tokenization is splitting text into smaller units called tokens. A sentence is a token of a paragraph, and a word is a token of a sentence, and the character is a token of the word. Using tokenization is a deep learning model such as RNN, LSTM, and GRU processing the text in the form of tokens. We have performed word-level tokenization for our project.

Removing Stopwords:
Stop words are a popular and common term in natural language processing. Removing stop words is performed to remove unnecessary and standard terms from text. One can think of removing such types of words may lead to misleading the model. However, this is not the case. Here, unnecessary means the terms (words) without which the text's sentiment or meaning is still achievable. We remove the preposition, conjunction, noun, pronoun, and other repeated and unusual words from the text without which the text's sentiment can be obtained.

Stemming:
In most of the text related models stemming is performed to get the root form of the word. Words can be in different forms based on the context of the sentence. A word can appear as a verb in a sentence, whereas in other sentences, it can appear as an adverb. Also, based on the sentence structure, a word can be in the present, past, or past participle form. If we consider including all these words instead of stemming these words into their root form as part of our model, then dictionary size will be huge, and also, the model will perform a massive amount of unnecessary calculation, which is an inefficient use of the memory. So, ultimately the model will take more time during the training process. To avoid all these, we perform stemming operations to convert each word into its root form.

Recurrent Neural Network (RNN):
Traditional feed forwarding networks don't have the ability to memorize anything over time. If the input and associated weights are considered as a constant, the output will also be constant. This behavior creates many issues in case of continuous input and when the model has to consider previous input to make predictions. To overcome this scenario, RNN is used rapidly. In RNN, inputs are provided as a sequence, and as it has a memory associated with it, it takes the output of the previous states as an input to the current state. The backpropagation is performed over the time in which it takes into account all the previous states weights associated with the current state. RNN faces two issues, namely: Vanishing gradient and the Exploding gradient. As the backpropagation is calculated over time, considerably in a large model, the gradient value eventually becomes a very little value as it is the multiplication of all the previous states' weights. This tiny gradient value ultimately affects the learning rate to grow very slowly. Thus, in most cases, we will never reach the local minimum or global minimum. In other cases, sometimes, during the training process, the model assigns some large weights without any consideration, which is irrelevant compared to the neighboring weights.

Long-Short-Term-Memory (LSTM):
To overcome the issue of vanishing gradient, LSTM is used. LSTM is just another type of RNN but with some modifications. To recall what model is seen a long time back, long term memory is used, and to recollect recent events; short term memory is used. The fundamental working of LSTM performed by four gates. The forget gate remembers the necessary parts from the previous states based on the current event and forgets unnecessary portions. The learn gate learns from the current input and the recent events. The remember gate takes the input from the forget

to get and learn the gate and remembers everything required for future reference of the model. The use gate maps the output from the forget gate and learns gate, considers all the factors, and provides the prediction of the current input. We used LSTM with 256 hidden nodes, and to avoid overfitting, dropping probability is used. The last layer is fully connected.

Results:-

Our ultimate goal of this project is to make secure cyberspace where anyone can feel safe and make communication through social media positively and effectively. The project is built, considering increasing hate spreading throughout the social media and extensive involvement of the young generation. Our model outperforms the traditional method of fighting against cyberbullying and also able to prevent users from doing this kind of activity on social platforms. By using Long-Term-Short-Memory, we can get an accuracy of 95%.

Conclusion:-

In this paper, we build a framework to detect cyberbullying and propose a mechanism to prevent it. To the most effective of our knowledge, none of the previous studies looks into how cyberbullying behavior is often prevented in social media. RNN with LSTM is employed to classify the model, and using Flask; the model is deployed online. The proposed system helps in removing any human moderator for social monitoring. Human moderators would come short just to monitor social networks because of the massive number of active users. As the deep learning model is used for detection, the performance is better than the traditional mechanism, and we also know that deep learning model is a mimic of the human brain that's why over time, the model will be more accurate and processing time will be much faster. However, this model may be extended with further future improvements.

Future Scope:-

Presently the model is implemented only on this blogging website; the next time, it can be implemented on different sites for monitoring. Now the model is only deployed on web applications; shortly, it can be implemented on Android apps and iOS also. The new model is limited to detection only on posts; in the future, we would implement it in different parts like comments and message texting. The accuracy of the model can be further improved with better architecture and through experiments with different hyperparameters.

Acknowledgement:-

The authors are thankful to the Department of Information Science and Engineering, Acharya Institute of Technology, Bangalore, India for extending their support during the study.

References:-

1. Mifta Sintaha, Moin Mostakim, "An Empirical Study and Analysis of the Machine Learning Algorithms Used in Detecting Cyberbullying in Social Media."
2. Hani Nurrahmi, Dade Mirjana, "Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility."
3. Kim D. Gorro, Mary Jane G. Sabellano, Ken Gorro, Christian Maderazo, Kris Capao, "Classification of Cyberbullying in Facebook Using Selenium and SVM."
4. Monirah A. Al-Ajlan, Mourad Ykhlef , "Optimized Twitter Cyberbullying Detection based on Deep Learning.",.
5. Sabina Tomkins, Lise Getoor, Yunfei Chen, Yi Zhang, "A Socio-linguistic Model for Cyberbullying Detection."
6. Vijay Banerjee,Jui Telavane, Pooja Gaikwad, and Pallavi Vartak, "Detection of Cyberbullying Using Deep Neural Network."
7. Djedjiga Mouheb, Masa Hilal Abushamleh, Maya Hilal Abushamleh, "Real-time Detection of Cyberbullying in Arabic Twitter Streams."
8. https://medium.com/@purnasaigudikandula/recurrent-neural-networks-and-lstm-explained-7f51c7f6bbb9.