A study of cyberbullying detection using Deep Learning and Machine Learning Techniques

Sagar Dalal[1], Prasad Pophalkar[2]

[1]Sagar Dalal MIT School of Engineering, Pune
[2]Prasad Pophalkar MIT School of Engineering, Pune
[3]Prof. Dr. Nitish Das , Dept. of Computer Science & Engineering, MIT School of Engineering, MIT ADT University, Pune, Maharashtra, India

Abstract - *The use of social media is widespread and it is ordinary for society of variable ages to have an account on social media platforms. Misuse of social media profiles has become quite easy for the same reason.. Cyberbullying or bullying through social media websites have been an ongoing issue that has been raised with the creation of social media. Since it has adverse effects on people's psychological behavior and mental health , timely detection and prevention of social media bullying is a very important factor in the world of the internet. Machine learning and Deep learning are the typical approaches adopted for cyberbully detection. In this research paper, we take a dataset containing tweets from the popular social media website Twitter to find texts that are possible cases of bullying. We create a hybrid bully detection system named Stacking Algorithm by merging three ML algorithms – K Nearest Neighbor (KNN), Support Vector Machine (SVM) and Random Forest (RF) to detect cyberbully texts in an accurate manner. The texts are checked and classified into three categories – Not Bullying, Racism and Sexism. We also include a section of Convolutional Neural Network (CNN) which identifies the bully texts accurately. A GUI is designed to display the accuracy percentage of the hybrid system as well as CNN . The accuracy measures of the two systems are evaluated, and it is determined that CNN produces a more precise estimate than the Stacking Algorithm.*

Key Words: Cyberbullying, Social Media, Machine Learning, Deep Learning

1. INTRODUCTION

Cyberbullying can be called bullying or harassment across digital media , mobile phones, laptops, and PCs are examples of such mediums.. Social media bullying refers to the bullying or harassment through websites like YouTube, Twitter, Facebook and Instagram. It involves sharing personal data regarding somebody across the web, posting hateful or abusive comments, accusations, threatening and hateful words, and so on . This issue is common among teenagers. However, with the speedy increase within the use of social media by older folks, social media harassment has become a typical downside among folks notwithstanding age. During a recent survey conducted, 47% young people, Children, kids, Youngsters have been receiving frightful comments in their social media profiles and 62% percent people are sent frightful personal messages. Despite the fact that there are methods for reporting bullying allegations, 91% of those who did so indicated that no action was taken.. Social media bullying victims bear mental state problems therefore it's necessary to develop a system that may discover bullying accurately and exactly. Many researchers are tired of this field principally victimizing machine learning and deep learning techniques. Random Forest, Support Vector Machine, Naive Bayes, and KNN are some of the most often used ML algorithms.

Similarly, we used Convolutional Neural Network (CNN) , a deep learning approach adopted to detect cyberbullying. In previous works related to the topic, several attempts have been made to improve the existing algorithms through different methods. One such method is introducing new features to the algorithm used. Personal details of a user in social media, their popularity, number of followers and following etc. are few examples of features. These features have been mixed and matched in different ways to enhance precision of the algorithm and to find the combination of features that produces best precision. Also, several studies were conducted to compare the performance of different ML algorithms and it could be concluded that SVM is the best technique out of all. Along with consideration of textual cyberbullying, visual cyberbullying was also studied. CNN is an algorithm suitable for finding bullying through images or videos. But very few attempts have been made to experiment in this field and there is still lots of scope to improve in this area. Other attempts mainly included bully detection in languages other than English such as Dutch, Bangla etc. and attempts to use the same algorithm in different languages.

In machine learning, there is a Training Dataset and Testing Dataset. In this research, a dataset is taken from the social media website Twitter. The data which is in an unstructured form is cleaned by correcting spelling mistakes, removing stop words and forming tokens of words. 80% of that data is used to train the algorithm while 20% data helps to test the algorithm. The output of this research categorizes data into three, namely, non-bully, racism and sexism. When extracting data from social media it may contain a mixture of words and characters such as special characters, words in different languages and alphanumeric characters. The data is mostly unstructured. This makes it difficult for traditional bully detection methods to detect bully texts precisely. Therefore, machine learning has emerged as the best approach to detect social media bullying as it is an automatic detection method. Hence, we propose a system which combines the top three ML algorithms – SVM, RF and NB to form a hybrid algorithm called Stacking algorithm. Also, an analysis is done using CNN to check the accuracy in predicting bully texts. The data is first fed into CNN. It produces a Confusion matrix which shows the accuracy percentage of texts in each of the three categories. It also displays a text box where you can manually enter any data from the dataset to see the category to which it belongs to. After CNN is run, Stacking Algorithm starts to run and it follows the same steps to display the number of tweets falling under the categories.

Hence, the aim of this research is to:

1. To study data from Twitter and provide accurate cyberbullying prediction, create a mix of algorithms that incorporates the top 3 machine learning algorithms.

2. To assess how accurate the hybrid system is in making predictions, compare it to the current CNN algorithm.

2. RELATED WORK

As per the findings of a study conducted by (Jang, et al., 2019) it's clear that CNN will follow 2 forms of approaches OnehotEncoder and Wordtovec for generation of vectors. The results of trials done by (Jang, et al., 2019) showed that wordtovec improves the system's overall accuracy when compared to onehotencoder. As a result, wordtovec is the most popular choice for this thesis.

According to the results of a similar study between CNN and KNN algorithms conducted by (Yin, et al., 2017), RNN surpasses CNN when it comes to a wide range of tasks, but CNN results square measure higher with key phrase detection and discussion based mostly sentences, which is precisely the type of work that will be performed during this thesis. As a result, the thesis may employ CNN rules to analyze activity sentiment.

Cyberbullying detection incorporates a speedily growing literature, despite the fact that research addressing bullying square measure derived back to the beginning of 2010. The extensive literature in this topic may be classified into different types: content-based detection, user-based detection, and network-based sensing.

2.1 Content-Based Detection

One of the first to combat cyberbullying on social media is, where a framework was designed to integrate Twitter streaming API for aggregating tweets and categorizing them given the content. Their work used elements of sentiment analysis and harassment identification. Tweets are classed as beneficial or harmful in the first part, then as positive containing bullying material, positive while not including bullying content, negative having bullying content, and negative while not bullying content. For categorization purposes, Nave mathematician was used, which resulted in a very high accuracy (70 percent). Another subsequent investigation used applied mathematics metrics, specifically (TFIDF) and (LDA), with topic models to uncover connections in documents. Hey did not, however, consider just applied mathematics metrics, but also extracted content possibilities such as: harmful words and pronouns. Alternative researchers continued to explore cyberbullying detection from a content-based standpoint, but they added new alternatives such as an emotion symbol and a hieroglyphic lexicon. Their method was put to the test by exploiting numerous learning algorithms, including Nave mathematician, SVM, and J forty eight. The best result was obtained with SVM, which achieved an accuracy of eighty one percent. Another analysis provided an example technique that organization members may use to monitor social networking sites and identify bullying instances. The subsequent strategy depended on capturing bullying phrases and keeping them in a massive quantity of data, therefore using Twitter API to catch tweets and comparing their content to the harassment content captured previously. Despite the potential revolutionary strategy in their work, this example system has yet to be implemented.

2.2 User-Based Detection

Many academics assumed that user data like age and the number of tweets sent may signal a person's capacity to harm others. The detection approach used user data such as the number of tweets sent, the number of followers, and the number of followers. Their combined options—user-based Associate in Nursing others—resulted in accurate forecasts with an accuracy of 85%. Similarly, they include a user's age as a feature in addition to a user's history as a feature. They believe that if a user has been bullied in the past, he will be more likely to be bullied again. They looked at the impact of introducing user options and discovered that it improves five-hitter recall.User-based choices were also introduced, where user gender and age were added to the set of features. The concept was that various genders speak different languages and people of different ages write in diverse styles. Furthermore, the user location was included as a completely new user feature.

2.3 Network-based Detection

The social organization of users is an interesting approach to detecting cyberbullying. Drawing the network structure and etymologizing alternatives from the graph are the first steps. They focused on deriving etymologies from social media alternatives.

Diagrams of connections The number of nodes indicates how big the community is, and the number of edges indicates how connected it is. Another study focusing on network-based possibilities has been discovered. They employed a graphical interface called (Gephi) to do their work.

The bullying postings were backed by the property. Then they looked into the role of the participants in the bullying, whether they were targets or attackers.

3. LITERATURE REVIEW

[1] Rui Zhao and Kezhi Mao proposed as follows :

They implemented a brand-new illustration learning approach that had been developed specifically to address this issue. The Semantic-Enhanced Marginalized Denoising AutoEncoder (smSDA) is a linguistics expansion of the popular stacked denoising autoencoder deep learning model. The linguistics extension includes linguistics dropout noise and scantness restrictions, with the linguistics dropout noise supporting domain information as well as the word embedding approach. This method can learn a strong and discriminative text illustration by utilizing the hidden feature structure of bullying data. This study aids in the understanding of denoising and autoencoding, and so proven to be beneficial for the more cost-effective display of data.

[2] ElahehRaisi and Bert Huang proposed as follows:

An unsteady supervised machine learning strategy for inferring user roles in harassment-based bullying as well as novel bullying word markers. The educational rule takes into account structure and concludes that users are more likely to bully and be exploited. The rule uses an outsized, unlabeled corpus of social media interactions to extract bullying roles of users and extra vocabulary indicators of bullying, and it also uses an outsized, unlabeled corpus of social media interactions to extract bullying roles of users and extra vocabulary indicators of bullying. Every social encounter is assessed to see if it is bullying supported by the model. The UN agency takes part and supports the language used, and it attempts to optimize the agreement between these projections, i.e. participant-vocabulary consistency (PVC).Through this paper the role of PVC is studied and also the data associated with the detection of the bullying roles of users is learned.

[3] P. Zhou proposed as follows:

An Attention-Based two-way Long STM Networks (Att-BLSTM) to capture the foremost vital linguistics data in an exceedingly large sentence. The experimental results on the SemEval-2010 relation classification task show that this technique outperforms most of the prevailing ways, with solely word vectors. This paper proposes a unique neural network Att- BLSTM for relation classification. This model doesn't utilize any options derived from lexical resources or natural language processing systems. The contribution of this paper is victimization BLSTM with AN attention mechanism, which might mechanically target the words that have a decisive result on classification, to capture the foremost vital se- prophetical data in an exceedingly sentence, while not victimizing further information and natural language processing systems. Through this paper the BLSTM paying attention to neural networks and options of BLSTM to classify the info a lot of accurately is studied.

[4] N. Srivastava proposed as follows:

Explains the dropout approach. Dropout increases neural network performance on supervised learning tasks in vision, speech recognition, document classification, and procedural biology, according to the report, with progressive outcomes on numerous benchmark data sources . The key plan is to haphazardly drop units (along with their connections) from the neural network throughout coaching. Throughout coaching, dropout samples from Associate in Nursing exponential range of various "thinned" networks are taken. At take a look at time, the result of averaging the predictions of these dilute networks will be therefore approximated by merely employing a singular untinned network that has lesser weights. This considerably reduces overfitting and offers major enhancements over alternative regularization ways. This paper helps to grasp the benefits and uses of dropout. the consequences of dropout will be studied.

[5] A. Conneau proposed as follows:

The fundamental plan of ConvNets is to contemplate feature extraction and classification united conjointly trained tasks. This paper presents a replacement design (VD-CNN) for a text process that operates directly at the character level and uses solely tiny convolutions and pooling operations. The paper shows that the effectuation of this model will increase with the deeper level, sign up to twenty nine convolutional layers, and report enhancements over the progressive on many public text classification tasks. ConvNets square measure chiefly custom-made for laptop vision owing to the integrative structure of a picture. Characters combine to form n-grams, stems, words, phrases, and sentences, among other things. This study explains how to use deep convolutional networks for text categorization and the advantages they have over RNN and LSTM.

[6] S. Bhoir proposed as follows:

A comparison of different word embedding methods, including Seamless bag of words, Skip grams, Glove(Global Vectors for word representation), and Hellinger PCA, was presented (Principal part Analysis). On entirely different criteria, the models are evaluated. Performance is measured in terms of the quantity of coaching data, a basic summary, and the relationship between context and targeted phrases, memory utilization, the supported predictor employed, and the influence of changes in spatiality. The process of word embedding converts text into numbers. This transformation has two important properties: spatiality reduction for cost-effective visualization and discourse similarity for communicative representations. Thus, this paper proves helpful to grasp the advantages of assorted word embedding models and also the comparison between them therefore on choosing the most effective model.

[7] E. Raisi proposed as follows:

The participant-vocabulary consistent model was presented as a debile supervised strategy for understanding the responsibilities of people on social media in harassment cyberbullying as well as the likelihood of language signals being used in such cyberbullying. Because the PVC's training process incorporates the structure of the communication network, it will find instances of apparent bullying as well as novel bullying indications. It assesses PVC on all social media platforms with each quantitative and analytical data collection. The weekly supervised method extrapolates from weak signs to find possible bullying episodes in the news. After that the discernible information from SMP's is formalized. There's a model designed that checks for the complete speech communication between the two persons victimization the designed score and victim score. This model needs no additional area on the far side for the storage of vectors and data. Thus, this paper helps in learning the PVC technique.

[8] H. Zeng proposed as follows:

To explore learning factors, a mental picture approach with four connected perspectives was used. Because of the neural standard, AlexNet is used in this investigation. In order to improve the coaching approach, we need to encourage the understanding of how the model parameters change from lower to greater accuracy. The two major obstacles to understanding the relationship between model parameters and performance, namely measurability and interpretability, have been overcome. The four reads that the system provides are area unit spec view, distinction distribution read, convolution operation read and performance comparison read. All the four views combined facilitate to grasp the insight of CNN clearly. numerous parameters and activation values between 2 CNN snapshots area unit evaluated supported the TFlearn framework. Because the coaching method of CNN results in an outsized variety of factors over time, this leads to remittent performance. This paper helps to look at the educational method of CNN.

[9] V. N. Kumar proposed as follows:

For proper learning, appropriate content illustration is critical. This study uses a naive mathematician as a classifier for content classification in email applications. It deals with the classification of spam terms after a message is received and processed using feature set extraction methods, where feature chances are obtained. For the precision problem, NB and SVM are examined. The message is simply categorized as cyberbullying in this report. By grouping massage, the denoised worth of each word is computed. The system is alerted by this system. It employs a word embedding approach to mechanically get the bullying character. Interface planning, training dataset, categorization, and analysis of twitter messages are some of the modules used in this work . Robust illustration and learning of text messages are crucial for a standardized detection system. The best methodology for the information extraction is internet primarily based mining technology.

[10] Andrew M. Dal and Quoc V. Le proposed as follows:

A paradigm for guided sequence learning victimization CNN and LSTM. Semi-supervised learning is a blend of supervised and unsupervised learning in which the unclassified information is used to see whether it may aid with the generalization of future supervised models. The article suggests that LSTM-RNN is more useful than CNN and RNN for the purpose of information training when using the proposed technique. The sequence autoencoder is employed here to reconstruct the input sequence itself i.e. the initial sequence. This paper uses LSTM thanks to bound edges like maintaining data ordering. This paper checks the semi-supervised technique on 5 benchmarks to see the results of victimization LSTM because of the coaching technique. This paper proves that CNN-LSTM is a higher technique than standard CNN and provides higher results than the previous ways for coaching unlabeled information.

[11] K. Duan proposed as follows:

Explains each one-versus-all and one-versus-one classifier in the SoftMax combination for multicategory classification. This study discusses how to expand the binary classification approach for multi-category classification efficiently . The paper conjointly explains that the majority common approaches to multi-category classification are unit binary-classifiers based mostly strategies like "one-versus-all" and "one-versus-one" that solves the multicategory classification downside. The one-versus-all methodology is sometimes enforced employing a "winner-takes-all" strategy. Whereas the one-versus-all methodology is sometimes enforced victimization liquid ecstasy wins vote strategy here the multicategory classification methodology is outlined victimization these 2 classifiers through a SoftMax operation. posteriori chances obtained from the mixture area unit accustomed do multicategory classification. This paper helps perceive the benefits of multicategory classification and conjointly strategies to implement that victimization 2 strategies thoroughly.

[12] Q. Li proposed as follows:

A novel technique to tweet sentiment categorization mistreatment SSWE and WTFM manufacture categories supported the weight theme and text negation and a replacement text classification methodology. The tactic here is evidenced to be higher than the SSWE and Nuclear Regulatory Commission techniques. During this the sentiment of tweets is polarized into three sorts. The paper suggests the SSWE word embedding formula for information illustration because it additionally will affect sentiment classification. The WTFM has a pair of options i.e. negation feature and also the tf.idf word consideration theme. Within the model here (SSWE + WTFM) the four options of WTFM concatenated with SSWE. The SSWE captures the linguistics and syntactical options and mistreats the initial n-gram polarity of tweets; it predicts a 2-dimensional vector (f0, f1).

[13] A. EI Adel Proposed as follows:

Deep Convolutional Neural Networks square measure used for the dropout and layer skipping. There's a key advantage: fast thanks to figure the feature mistreatment quick beta wave remodel. The purpose intelligent dropout methodology.is based on a unit is potency and not indiscriminately elect.it is potential to classify the image mistreatment economical unit of earlier layer and skip the all hidden layer from the output layer. This paper proves that the FWT is the best or it's extracting the feature of the input image.

[14] S. Zhai proposed as follows:

Investigated and forecasted the context of search-based online advertising. We utilize a repeating neural network to translate each query and add to real valued vectors so that the relevance of a specific combination may be easily determined. We assign a distinct attention value to each word position based on its purpose and relevance. During this, the vector output of a sequence is generated by a weighted ad of the RNN's hidden state at each word according to their attention score. This research demonstrates how the RNN allows North American nations to model word sequences, which are demonstrated to be very important in properly capturing the meanings of a series.

[15] I. Raid proposed as follows:

Explained development and advancement of technology additionally to transportation a positive impact and conjointly introduced new issues once used unsuitably. Typically this can be often observed as law-breaking. one in every law-breaking is being spirited at the instant is cyberbullying. Social media is one in every of the for the event of cyberbullying .The analysis was administered victimization method of data processing There are various processes such as data collection, preprocessing, TF-IDF, weighting, information validation, and classification using a naïve mathematical classifier. This paper proves that TF-IDF weighted and validates information victimization cross validation and so will classification.

[16] K. Sahay proposed as follows:

Explains how online bullying and aggressiveness towards social media users has exploded in recent years. It affects about half of all young netizens, and the ongoing application of insult detection victimization machine learning and language communication processes has a terrible recall rate. The assessment experimenting with completely different work methods results in a robust methodology for extracting text, user experimenting with completely different work methods results in a strong methodology for extracting text, user adding certain ways to spot and classify bullying within the text by analyzing and network-based attributes learning the properties of cyberbullying and aggressors and what features distinguish them for normal users the information processing and machine learning a This study demonstrates coaching in ML models using supervised learning.

4. METHODOLOGY

4.1 Data Description

The section can discuss the steps taken whereas developing the system to observe cyber bullying for twitter. This section also will discuss varied tools and packages used for developing the system.

The algorithm that was utilized to create the paper:

· Stacking ensemble (K-Nearest Neighbor + Support Vector Machine + Random Forest)

· Convolution neural Network (CNN).

The study may utilize a CNN methodology and a stacking ensemble to find any harassment emotion gift in any tweet from Twitter.

Stacking Ensemble(Rocca, 2019):

Because it involves many algorithms acting on the same dataset, stacking might be considered a hybrid style of machine learning algorithmic rule. Stacking is made up of two levels of classifiers. the meta classifier as well as the bottom level classifiers On a computer file, the lowest level classifiers are permitted to function in parallel and one by one. The bottom level classifier's findings are sent into the meta level classifier as input. The meta level, or ultimate classifier, that supports the received input predicts the final outcomes. In the case of a stacking algorithmic rule, many classifiers care about one input, reinforcing the model's correctness. Any disadvantages caused by one model's output may be overcome by the results of another model. Finally, the final level classifiers eliminate all prior mistakes and improve overall accuracy.

CNN (Convolution Neural Network)(SHARMA, 2018):

CNN is a type of neural network that is built on numerous filters and rules. Learnable filters are the type of filters that can be learned. The filters, as well as the learnable filters gift, are of the perfect size. These filters are applied to every input file. While the information is being processed, each of the filters operates on it. Strides is a filter operation that computes the real number between the weights of the filters and the patch out from the receptive field.

A convolution neural network has the following layers:

· Input layer

· Convolution layer

· Activation function layer

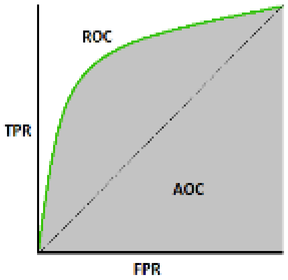
· Pool layer

· Fully connected layer

Proposed Evaluation:

A confusion matrix, AUC-ROC Curve, precision, recall, and f1-score worth may be included in the outcome for each formula. These calculations can facilitate a correct analysis of results obtained by every formula well.

AUC-ROC Curve:



AUC-ROC Curve conjointly referred to as Separability curve. It represents the degree of disjuncture between the given knowledge points. The lower its worth the higher is the classification.

Confusion Matrix:

Once the data set is entered into a formula, it will handle the count of False Positives, False Negatives, True Positives, and True Negatives.

All these inputs can facilitate Pine Tree State to use the dataset to do a comparative analysis of algorithms. These values can facilitate the investigation of which algorithmic program performs well during which department. For instance, that algorithmic program has higher accuracy, that algorithmic program consists of low false positive and false negative. we'll conjointly see if or not hybrid algorithms will have higher results than the opposite algorithms. We'll verify whether or not victimization spark framework over Hadoop generates any performance increase of the process.

5. CONCLUSIONS

In this analysis, we will compare the algorithms used by hybrid assembly and neural network algorithms in order to determine which type of machine learning strategy is most suited to solving the problem of cyberbullying on social media. We've noticed that CNN-based neural networks operate better or perform better than other neural networks. According to the literature study, several supervised machine learning techniques are used to create sub-classifiers for stacking methods. We can infer from the data analysis and comparison research done for each algorithm listed in the result section that stacking-based machine learning techniques are superior to neural network-based approaches for detecting harassment or bullying on social media platforms. The accuracy obtained for stacking is around 83 percent, whereas the accuracy acquired for the Convolution Neural Network is nearly 80 percent.

REFERENCES

[1] "Text classification using convolutional 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). "Modeling 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)

[18] S. Salawu, Y. He, and J. Lumsden, "Approaches to Automated Detection of Cyberbullying : A Survey," vol. 3045, no. c, pp. 1–20, 2017.

[19] T. Wu, S. Liu, J. Zhang, and Y. Xiang, "Twitter spam detection based on deep learning," Proc. Australas. Comput. Sci. Week Multiconference - ACSW "17, pp. 1–8, 2017.