A Machine Learning Technique to Identifying the Cyber bullying in Social Media

Nissankara Taraka Ramchand, Alekhya Narala, Rambabu Gouru, Vamshi Gangaraboina
Dr. Deena Babu Mandru
Department of Information Technology, Malla Reddy Engineering College, Secunderabad, Telangana-500100

Abstract

Cyberbullying is a major problem encountered on internet that affects teenagers and adults. It has lead to mishappenings like suicide and depression. Regulation of content on Social media platforms has become a growing need. The following study uses data from two different forms of cyber bullying, hate speech tweets from Twittter and comments based on personal attacks from Wikipedia forums to build a model based on detection of Cyberbullying in text data using Natural Language Processing (NLP) and Machine learning. Three methods for Feature extraction and four classifiers are studied to outline the best approach. For Tweet data the model provides accuracies above 90% and for Wikipedia data it gives accuracies above 80%.

Keywords: Cyber bullying, Social Media, NLP, Semi-supervised learning, feature extraction.

I. INTRODUCTION

Millions of youthful individuals spend their time on social organizing, and the sharing of data online. Social systems have the capacity to communicate and to share data with anybody, at any time, and within the number of individuals at the same time. There are over 3 billion social media clients around the world. Agreeing to the NCPC, cyberbullying is online where portable phones, video diversion apps, or any other way to send or send content, photographs, or recordings purposely harm or humiliate another individual. Cyberbullying can happen at any time all day, weekand you'll reach anybody anyplace through the web. Content, photographs, or recordings of cyberbullying may be posted in an undisclosed way. It can be troublesome, and now and then inconceivable, to track down the source of this post. It was too in comprehensible to induce freed of these messages after ward. A few social media stages such as Twitter, Instagram, Facebook, YouTube,

Snapchat, Skype, and Wikipedia are the foremost common bullying locales on the web. A few of the social organizing locales, such as Facebook, and the arrangement of direction on the prevention of bullying. It encompasses, an extraordinary segment that clarifies how to report cyber-bullying and to anticipate any blocking of the client. On Instagram, when somebody offers photographs and recordings made by the client to be awkward, so the client can screen or piece them. Clients can too report a infringement of our Community and make Suggestions to the app. As the social way of life surpasses the physical obstruction of human interaction and contains unregulated contact with outsiders, it is fundamental to analyze and think about the setting of cyberbullying. Cyberbullying makes the casualty feel that he is being assaulted all over as the web is fair a tap absent. It can have mental, physical, and enthusiastic impacts on the casualty.

II. LITERATURE SURVEY

Parcel of investigate has been done to discover conceivable arrangements to identify Cyber bullying in social organizing locales. Hsien [1] utilized an approach utilizing watch word coordinating, supposition mining,and social arrange examination andgot an accuracy of 0.79 and a review of 0.71 from datasets from four websites. Patxi Gal´an-Garc´ıa et al. [2] proposed a theory that a troll on a social organizing locales beneath a fake profile continuously includes a genuine profile to check how others see the phony profile. They proposed a Machine learning approach to decide on such profiles. The distinguishing proof handle examined a few profiles that have a few close relations to them. The strategy utilized was to choose profiles to consider, procure data of tweets, select highlights from profiles, and utilize ML to discover the creator of tweets. 1900 tweets were utilized ,having a place for 19 distinctive shapes. It had a precision of 68% for recognizing the creator. Afterward, it was utilized in a Case Ponder in a school in Spain where out of a few suspected understudies for Cyberbullying the genuine proprietor of a profile had to be found and the strategy worked within the case. The

taking after strategy still has a few inadequacies. For illustration a case where trolling account doesn't havea genuine account,to trick such frameworks or specialists who can alter composing styles and practices so that no designs are found. For changing composing styles,more productive calculations will be required.

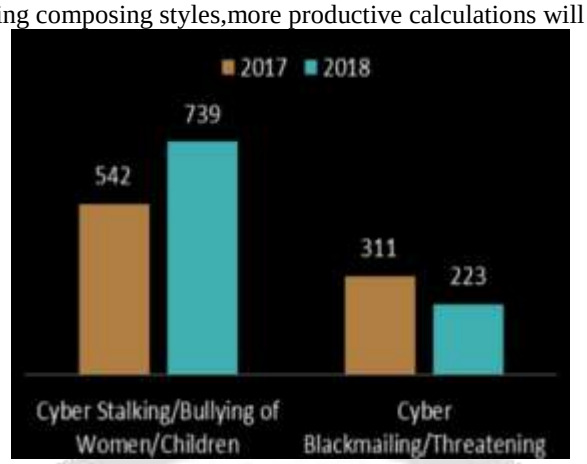

Fig 1. Literature Survey

III. PROPOSED METHODOLOGY

Below fig-2 represents the proposed system architecture.It mainly contains loading the original dataset, data preprocessing, feature extraction and classification stages.

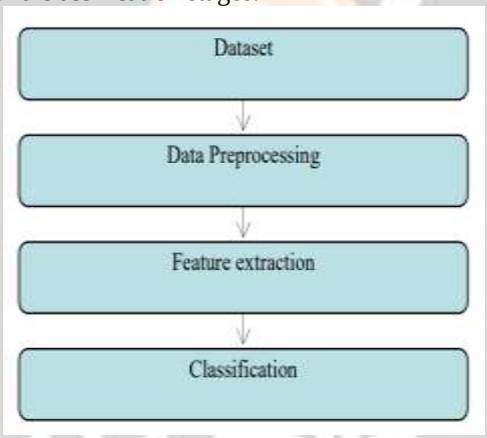

Fig 2. Architecture of proposed Methodology

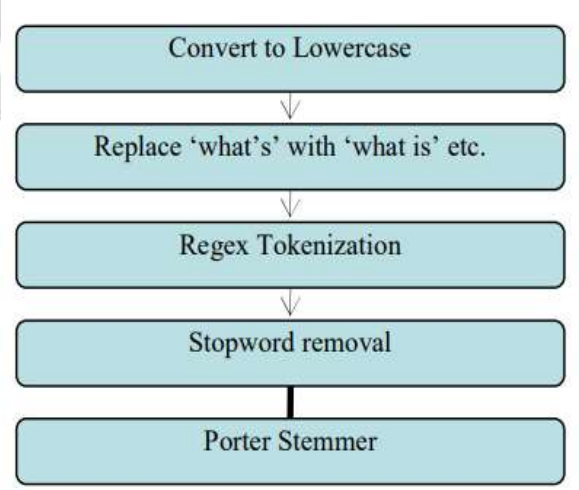

Fig 3. Data Processing Pipeline

A. Machine Learning Algorithms

In this section, we discussed the basic mechanisms of several machine learning algorithms. We presented Decision Tree, Naive Bayes, Random Forest and Support Vector Machine in each subsection.

i. Decision Tree

The decision tree classifier can be used in both classification and regression. It can help represent the decision as well as make a decision. The decision tree is a treelike structure where each internal node represents a condition, and each leaf node represents a decision. A classification tree returns the class where the target falls. A regression tree yields the predicted value for an addressed input.

$$\mathbf{E(S) = -P(+) \log(p)(+) – P(-) \log(p)(-)}$$

Here, P(+) is the probability of positive class
P(-) is the probability of negative class

$$\mathbf{Information\ Gain = E(Y) – E(Y \mid X)}$$

ii. Naive Bayes

Naive Bayes is an efficient machine learning algorithm based on Bayes theorem. The algorithm predicts depending on the probability of an object. The binary and multi-class classification problems can be quickly solved using this technique. Based on Bayes' Theorem it finds the probability of an event occurring given the probability of another event that has already occurred as follows:

$$\mathbf{P(y|X) = P(X|y) \times P(y)\ X\ (1)}$$

Here, where, the class variable is denoted by y and X is a dependent feature vector of length n as **X = x1, x2, x3...xn.**

ii. Random Forest

Random Forest classifier is consists of multiple decision tree classifiers. Each tree gives a class prediction individually. The maximum number of the predicted class is our final result. This classifier is a supervised learning model which provides accurate result because several decision trees are merged to make the outcome. Instead of relying on one decision tree, the random forest takes the prediction from each generated tree and based on the majority votes of predictions, and it decides the final output. For example, if we have two classes namely A and B and the most of the decision tree predict the class label B of any instance, then RF will decides the class label B as follows: f(x) = majority vote of all tree as B

$$\text{Entropy} = \Sigma\ (–pi) * \log(pi)$$

Here, pi represents the relative frequency of the class you are observing in the dataset and c represents the number of classes.

iii. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be applied in both classification and regression alike a decision tree. It can distinguish the classes uniquely in n-dimensional space. Thus, SVM produces a more accurate result than other algorithms in less time. In practice, SVM constructs set a of hyperplanes in a infinite-dimensional space and SVM is implemented with kernel which transforms an input data space into the required form. For example, Linear Kernel uses the normal dot product of any two instances as follows:

$$K(x, xi) = \text{sum}(x * xi)$$

iv. K-Nearest Neighbor

K-NN is an example of a supervised machine learning algorithm. This algorithm makes use of similarity in observations or samples to make predictions for new ones. The assumption is that the more similar samples are, the more likely they belong to the same category or class. The K in the K-NN represents the number of nearest neighbors for which the decision of whether or not the new sample belongs to the same class. In this work, there are two main categories for which group of neighboring samples can be classified, namely:

$$\mathbf{h(x) = mode(\{y'':(x'',y'') \in Sx\})}$$

Here, mode(·) means to select the label of the highest occurrence

v. Extreme Learning Machine Algorithm

The extreme learning machine (ELM) is widely used in batch learning, sequential learning, and incremental learning because of its fast and efficient learning speed, fast convergence, good generalization ability, and ease of implementation. With the development of the traditional ELM, lots of improved ELM algorithms have been proposed; meanwhile the scope of implementing the ELM has been further expanded from supervised learning to semi-supervised learning and unsupervised learning. However, due to its memory-residency, and high space and time complexity, the traditional ELM is not able to train big data fast and efficiently. Optimization strategies have been employed for the traditional ELM to solve this problem. In this chapter, we will first review ELM theories and some important variants, and then describe parallel ELM algorithms based on MapReduce and Spark in detail. Lastly, we show some practical applications of the ELM for big data.

$$Fl(x) = \Sigma\ \beta g(x)$$

Here, L is a number of hidden units. B is the weight wector between th hidden layer and output. G is the activation function. X is an input vector.

IV. RESULT ANALYSIS

Fig 4. Algorithm Implementation

In above Figure we select each algorithm and click on 'Submit' button to train model and we will get accuracy for each algorithm. We have to repeat this step whenever first time we start the server or upon adding new bulling messages.

Fig. 5 Algorithm Accuracies

In above Figure We implemented SVM and got an accuracy of 95 percent. Check the below table to know more about the remaining algorithms accuracies.

Table-1: Accuracy Comparison Between Algorithms

Classifier	Average Accuracy
SVM Algorithm	86.36363
Naive Bayes Algorithm	95,.45454
Random Forest	54.5545
Decision Tree	86.3636
KNN Algorithm	59.09
Extreme Learning Machine	98.50

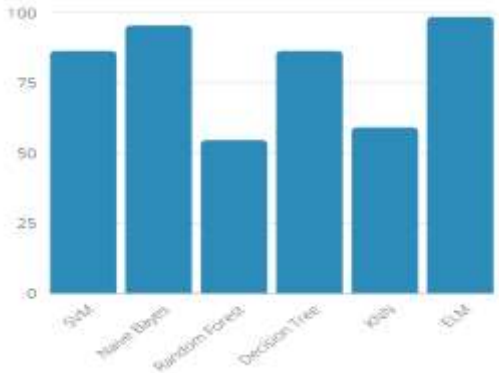

Fig. 6: Graphical representation of Accuracies

V. CONCLUSION

Cyberbullying across the internet is dangerous and leads to mishappenings like suicides, depression, etc., and therefore there is a need to control its spread. Therefore cyber bullying detection is vital on social media platforms. With availability of more data and better classified user information for various other forms of cyber attacks. Cyberbullying detection can be used on social media websites to ban users trying to take part in such activity In this paper we proposed an architecture for detection of cyber bullying to combat the situation.We discussed the architecture for two types of data: Hate speech Data on Twitter and Personal attacks on Wikipedia.For Hate speech Natural Language Processing techniques proved effective with accuracies of over 90 percent using basic Machine learning algorithms because tweets containing Hate speech consisted of profanity which made it easily detectable.Due to this it gives better results with BoW and Tf-Idf models rather than Word2Vec models However, Personal attacks were difficult to detect through the same model because the comments generally did not use any common sentiment that could be learned however the three feature selection methods performed similarly.Word2Vec models that use context of features proved effective in both datasets giving similar results in comparatively less features when combined with Multi Layered Perceptrons.

REFERENCES

[1] I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, “Towards the detection of cyberbullying based on social network mining techniques,” in Proceedings of 4th International Conference on Behavioral, Economic, and Socio-Cultural Computing, BESC 2017, 2017, vol. 2018-January, DOI: 10.1109/BESC.2017.8256403.

[2] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, “Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying,” 2014, doi: 10.1007/978-3-319-01854-6_43.

[3] A. Mangaonkar, A. Hayrapetian, and R. Raje, “Collaborative detection of cyberbullying behavior in Twitter data,” 2015, DOI: 10.1109/EIT.2015.7293405.

[4] R. Zhao, A. Zhou, and K. Mao, “Automatic detection of cyberbullying on social networks based on bullying features,” 2016, DOI: 10.1145/2833312.2849567.

[5] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, “Detection of Cyberbullying Using Deep Neural Network,” 2019, DOI: 10.1109/ICACCS.2019.8728378.

[6] K. Reynolds, A. Kontostathis, and L. Edwards, “Using machine learning to detect cyberbullying,” 2011, DOI: 10.1109/ICMLA.2011.152.

[7] J. Yadav, D. Kumar, and D. Chauhan, “Cyberbullying Detection using Pre-Trained BERT Model,” 2020, DOI: 10.1109/ICESC48915.2020.9155700.

[8] M. Dadvar and K. Eckert, “Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study,” arXiv. 2018.

[9] S. Agrawal and A. Awekar, “Deep learning for detecting cyberbullying across multiple social media platforms,” arXiv. 2018.

[10] Y. N. Silva, C. Rich, and D. Hall, “BullyBlocker: Towards the identification of cyberbullying in social networking sites,” 2016, DOI: 10.1109/ASONAM.2016.7752420.

[11] Z. Waseem and D. Hovy, “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter,” 2016, DOI: 10.18653/v1/n16-2013.