ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 1, 2024

DETECTION OF CYBERBULLYING ON SOCIAL MEDIA USING MACHINE LEARNING

D.Suma, Professor, Department Of CS SICET, Hyderabad
Gundlapally Varshitha,Basa Sairaj,Rangineni Deekshith Rao,Kalva Chennakeshava Reddy,Gollavelli Naga Surya
UG Student, Department Of CS, SICET, Hyderabad

ABSTRACT –
With the increase in social media users, cyber bullying has become a type of bullying done th rough electronic messages. Chats provide a rich environment that bullies can use to attack the ir victims. Considering the effects of cyberbullying on its victims, appropriate measures need to be taken to detect and prevent cyberbullying. Machine learning helps identify bullies' lang uage patterns, thereby creating patterns to detect cyberbullying. This article introduces a mac hine learning approach for the detection and prevention of cyberbullying. Many classification systems are used to educate and define bullying. Evaluation of the proposed method on the c yberbullying dataset shows that neural networks perform better with 92.8% accuracy and 90. 3 SVM. Additionally, the neural network performed better than other similar processes on the same data. Machine learning; Neural Networks

I.Introduction

With the increase in the number of media users, new methods of bullying have emerged. The second term is defined as the bad behavior or misbehavior of a person or group of people wh o communicate frequently from time to time towards victims who cannot easily protect thems elves [1]. Bullying is always a part of society. With the birth of the internet, it was only a mat ter of time before bullies took advantage of this new medium. By using services such as emai l and instant messaging, bullies can carry out their malicious behavior anonymously and awa y from their targets. According to the Cambridge Dictionary, the word cyberbullying is defin ed as the act of using the Internet to harm or threaten others, especially by sending negative m essages. The main difference between cyberbullying and traditional bullying is the impact on the victim. While traditional bullying can cause physical damage as well as mental and emoti onal damage, cyberbullying is all about emotional and mental damage. Work to detect and pr event it. An effective way to learn from data and create patterns that classify the correct beha viors is through machine learning. Machine learning helps identify bullies' language patterns and thus creates a framework for identifying cyberbullying. Therefore, the main task of this p aper is to propose a monitoring machine learning to detect and prevent cyberbullying. The pla n is measured through a cyberbullying dataset from Kaggle collected and tagged by authors K elly Reynolds and others. In his articles

[2]. Comparison of the performance of TFIDF and sentiment analysis feature extraction meth od, SVM and neural network classifiers. Additionally, experiments were conducted on differe nt ngram language models. 2 grams, 3 grams and 4 grams were taken into account when eval uating the design of the isolated product. Finally, we evaluate our plan against previous studie s using the same data. Various activities are presented in Part II. Section III describes the pro

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 1, 2024

posed method. Section 4 presents the experimental results and evaluation of the proposed method. Finally, Section 5 concludes the article.

Related Activities

There are many reporting systems that can detect cyberbullying with high accuracy. The first is from author Nandhini et al. [3] They implemented a model using Naive Bayesian machine learning, thanks to their work, they achieved 91% accuracy and obtained the dataset from MySpace.com, then reported the other model [4] Naive Bayes classifier and genetic operation (FuzGen), 87% attained righteousness. Another approach by Romsaiyud et al. [5] They developed a Naive Bayes class to extract words and look for shared patterns with this method, achieving 95.79% accuracy on Slashdot, Congressgate and MySpace datasets. But they have a problem, that is, the group process cannot run in parallel. Also in the scheme of Bunchanan et al. [6] They use the Tank War game discussion to obtain and classify the dataset, then compare it to simple classifications that characteristically use emotion theory, and their results do not compare well to the results of manual classification. Additionally, Issa et al. [7] After receiving datasets from Kaggle, they proposed a method, using two names: Naive Bayes and SVM. The average accuracy of the Naive Bayes classifier is 92.81%, while the average accuracy of the multcore SVM is 97.11%, but they did not mention the size of the training or testing data, so the results may not be reliable. Another approach by Dinakar et al. [8] To describe bullying related to (1) sexuality(2) race and culture, and (3) intelligence; They get information from the course on YouTube. After using SVM and Naive Bayes classifier, SVM achieved 66% accuracyNaive Bayes 63%. Continuing the discussion, DiCapua et al. [9] proposed a new method to detect cyberbullying using an unsupervised method. 67% recall of GHSOM was used on Twitter, achieving an accuracy of 60%, an accuracy of 69%, and a recall of 94%. Additionally, Haidar et al. [10] proposed a model to identify cyberbullying but in Arabic they used Naive Bayes and achieved an accuracy of 90.85% and SVM achieved an accuracy of 94.1% but also had a negative They have it and it's all in Arabic. A method proposed by Zhang et al. [11] used in their paper a novel speecbased convolutional neural network (PCNN) that reduces the problem of noise and data interference and thus overcomes class mismatch. 1313 comments came from Twitter and 13,000 comments from formspring.me. Its accuracy is not calculated due to the inconsistency of the Twitter dataset. While 56% precision, 78% recall and 96% accuracy are achieved, when high precision is achieved, their data is not equal and therefore gives erroneous results, which is reflected in the 56% score. Nobata et al. [12] recently reported the use of language exploitation, they used a framework called Vowpal wabbit for classification, they also developed a classification inspection method with NLP features that outperforms the deep learning method using summary of collected data, F.A score of up to 0.817 was reported on Yahoo News and Finance. [13] proposed a specific framework for cyberbullying, created a list of insults using word embeddings and focused on bullying, used SVM as the main distributor and achieved a response rate of 79.4%. Later, another method was proposed by Parime et al. [14] obtained datasets from MySpace and manually collected them and classified them using the SVM classifier. Additionally, Chen et al. [15] reported that new features, called morphosyntactic features, were extracted and vector machines were used as classification, achieving 77.9% accuracy and 77.8% recall. Additionally, Ting et al. [16] proposed an SNMbased method in which they collect data from relationships and then use SNA measurements and reasoning as features. Seven tests were performed and accuracy increased to 97% and recall to 71%. Additionally, Harsh Dani et al. [17] introduced a new framework called SICD and used K

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 1, 2024

NN for classification. They eventually achieved an F1 score of 0.6105 and an AUC score of 0 .7539. Dadwal et al. [18] [19] [20] [21] used WEKA to build support vector machine classifie rs on primary and secondary data and their data were collected from Myspace. They got 43% accuracy and 16% returns and didn't state that fact. The only difference between the two pape rs is that they use gender data for classification in the second form. Additionally, their second article collected 4626 comments from 3858 different users. These comments were labeled bu llying (9.7%) and not bullying (inteannotator consensus 93%). They used an SVM classifier a nd achieved up to 78% accuracyand up to 55% recall. Finally, in their third article, they used 3 models for data collected from YouTube tutorials: MultiCriteria Evaluation System (MCES), Machine Learning: (Naive Bayes Classifier, Decision Tree, SVM), mixed methods. MCES achieved an accuracy of 72%, while Naive Bayes achieved the highest score of the three at 66 %. Turning to other authors, Potha et al. [22] also used the SVM method and achieved 49.8% accuracy. Chavan et al. [23] used two types of classification: logistic regression and support vector machine. Logistic regression achieved 73.76 accuracy, 60% recall, and 64.4% precisio n. They obtained data from Kaggle, achieving 77.65% precision, 58% recall, and 70% precisi on for the support vector machine. ConceptAs shown in Figure 1, the plan has three main stag es: advancement, extraction and classification steps. In the first step, we clean the data by rem oving noise and unnecessary text. The pre-processing steps are as follows:

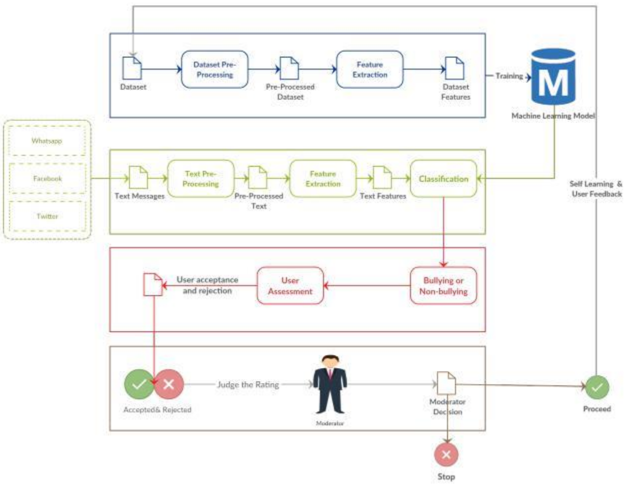

Fig. 1. Proposed Approach

The second step of the proposed model is the extraction step. In this step, data files are conve rted into a format suitable for machine learning algorithms. First, we use TFIDF [25] to extra ct features of the input data and put them into custom lists. The main idea of TFIDF is that it gets the weight of a word relative to a document or sentence by processing the text. In additio n to TFIDF, we also use the concept of semantic analysis [26] to extract sentences from sente nces and add them as features to the list of TFIDF features. The polarity of a sentence refers t o the classification of the sentence as positive or negative. To this end, we extract polarity usi ng the Text Blob library [27], which is a prelearned model for video analysis. In addition to u sing TFIDF and polarity inference for feature extraction, the method also uses NGram [28] to identify different word combinations during pattern evaluation. We specifically use 2 Grams, 3 Grams and 4 Grams. That is, it is used in the prediction phase. We use two types of classifi cation, SVM (Support Vector Machine) and Neural Network. A neural network has three layers: input layer, hidden layer and output layer. There are 128 nodes in the input layer. There are 64 neurons in the hidden layer. The output method is Boolean output. These criteria include accuracy, precision, recall, and f-score. They are calculated according to the following equation:

IV. EXPERIMENTAL RESULTS

This section describes the experimental results on the proposed approach. We evaluate the proposed approach on the cyberbullying dataset from kaggle. In the following we describes the Data and the results.

A. Data Description

We have used cyberbullying dataset from Kaggle which was collected and labeled by the authors Kelly Reynolds et al. in their paper [2]. This dataset contains in general 12773 conversations messages collected from Formspring.

cyberbullying or not. The annotation classes were unbalanced distributed such that 1038 question-answering instances out of 12773 belongs to the class cyberbullying, while 11735 belongs to the other class. First, to remedy the data unbalancing, we take the same number instances of both classes to measure the accuracy. We also removed from the data big size conversations and remove the noisy data. We ended up with total 1608 instance conversations where 804 instances belongs to each class. Table I summarizes the statistics of dataset.

TABLE I. STATISTICS OF THE DATASET

Total number of Conversations	1608
Number of cyberbullying	804
Number of non-Cyberbullying	804
Number of distinct words	5628
Number of token	48843
Maximum Conversation size	773 Characters
Minimum Conversation size	59 Characters

Results

After preprocessing the dataset, we follow the same step presented in Section III to extract the features. We then split the dataset into ratios (0.8,0.2) for train and test. Accuracy, recall and precision, and f-score are taken as a performance measure to evaluate the classifiers. We apply SVM as well as Neural Network (NN) as they are among the best perfor-mance classifiers in the literature. We run several experiments on different n-gram language model. In Particular, we take into consideration 2-gram, 3-gram, and 4-gram during the evaluation of the model produced by the classifiers. Table II summarizes the accuracy of both SVM and NN. The SVM classifier achieved the highest percentage using 4-Gram with accuracy 90.3% while the NN achieved highest accuracy using 3-Gram with accuracy 92.8%. It is found that the average accuracy of all n-gram models of NN achieves 91.76%, while the average accuracy of all n-gram models of SVM achieves 89.87%. Fig. 2 depicts the accuracy results of both classifiers.

V. CONCLUSION

In this paper, we proposed an approach to detect cyberbullying using machine learning techniques. We evaluated our model on two SVM and Neural Network classifiers and used TFIDF and sentiment analysis algorithms for feature extraction. The classifications were evaluated on different n-gram language models. We achieved 92.8% accuracy using a 3-gram neural network and 90.3% accuracy using a 4-gram SVM using both TFIDF and sentiment analysis. We found that our neural network performed better than the SVM classifier as it also achieves an average f-score of 91.9% while the SVM achieves an average f-score of 89.8%. We further compared our work with other related work that used the same dataset and found that our neural network outperformed their classifiers in terms of accuracy and f-score. By achieving this accuracy, our work will definitely improve cyberbullying detection to help people use social media safely. However, cyberbullying pattern detection is limited by the size of the training data. Thus, more cyberbullying data is needed to improve performance. Thus, deep learning techniques will be suitable in larger data as they are proven to outperform machine learning approaches over larger data sizes.

REFERENCE

[1] Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell and Neil Tippett. Cyberbullying: Its nature and impact on secondary school students. Journal of Child Psychology and Psychiatry, 49(4):376–385, 2008.

[2] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In 2011 10th International Machine Learning and Applications Conference and Workshops, Volume 2, Pages 241-244. IEEE, 2011.

[3] B Nandhini and JI Sheeba. Detection and classification of cyberbullying using an information retrieval algorithm. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), p 20. ACM, 2015.

[4] B Sri Nandhini and JI Sheeba. Online detection of social media bullying using intelligence techniques. Proceedings of Computer Science, 45:485–492, 2015.

[5] Walisa Romsaiyud, Kodchakorn na Nakornphanom, Pimpaka Prasert-silp, Piyaporn Nurarak and Pirom Konglerd. Automated cyberbullying detection using appearance pattern clustering. In Knowledge and Smart Technology (KST), 2017 9th International Conference on, pages 242– 247. IEEE, 2017.

[6] Shane Murnion, William J Buchanan, Adrian Smales and Gordon Russell. Machine learning and semantic analysis of in-game chat for cyberbullying. Computers and Security, 76: 197–213, 2018.

[7] Sani Muhamad Isa, Livia Ashianti et al. Classifying cyberbullying using text mining. In Information and Computational Sciences (ICICoS), 2017 1st International Conference on, pages 241–246. IEEE, 2017.

[8] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense for detecting, preventing and mitigating cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3):18, 2012.

[9] Michele Di Capua, Emanuel Di Nardo and Alfredo Petrosino. Uncontrolled detection of cyberbullying on social networks. In Pattern Recognition (ICPR), 2016 23rd International Conference on, pages 432–437. IEEE, 2016.

[10] Batoul Haidar, Maroun Chamoun and Ahmed Serhrouchni. A multilingual cyberbullying detection system: Arabic content detection using machine learning. Advances in Science, Technology and Engineering Systems Journal, 2(6):275–284, 2017.

[11] Xiang Zhang, Jonathan Tong, Nishant Vishwamitra, Elizabeth Whit-taker, Joseph P Mazer, Robin Kowalski, Hongxin Hu, Feng Luo, Jamie Macbeth, and Edward Dillon. Cyberbullying detection using an utterance-based convolutional neural network. In 2016, the 15 IEEE International Conference on Machine Learning and Applications (ICMLA), pages 740-745. IEEE, 2016.

[12] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Detecting offensive language in online user content. In Proceedings of the 25th International Conference on the World Wide Web, pages 145–153. World Wide Web International Conference Steering Committee, 2016.

[13] Rui Zhao, Anna Zhou and Kezhi Mao. Automatic detection of cyberbullying in social networks based on bullying features. In Proceedings of the 17th International Conference on Distributed Computing and Networking, page 43. ACM, 2016.

[14] Sourabh Parime and Vaibhav Suri. Cyberbullying detection and prevention: a data mining and psychological perspective. In Circuit, Power and Computing Technologies (ICCPCT), 2014 International Conference on, pages 1541–1547. IEEE, 2014.

[15] Ying Chen, Yilu Zhou, Sencun Zhu and Heng Xu. Detecting offensive language in social media to protect adolescents' online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on Social Computing and 2012 International Conference on Social Networks (SocialCom), pages 71–80. IEEE, 2012.

[16] I-Hsien Ting, Wun Sheng Liou, Dario Liberona, Shyue-Liang Wang, and Giovanny Mauricio Tarazona Bermudez. Towards cyberbullying detection based on social network mining techniques. In Behavioral, Economic, Socio-cultural Computing (BESC), 2017 International Conference on, pages 1–2. IEEE, 2017.

[17] Harsh Dani, Jundong Li and Huan Liu. Sentiment-based detection of social media cyberbullying. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 52–67. Springer, 2017.

[18] Maral Dadvar and Franciska De Jong. Detecting cyberbullying: a step towards a safer internet yard. In Proceedings of the 21st International Conference on World Wide Web, pp. 121–126. ACM, 2012.

[19] Maral Dadvar, de FMG Jong, Roeland Ordelman and Dolf Trieschnigg. Improved detection of cyberbullying using gender information. In Proceedings of the 1welfth Dutch-Belgian Information Retrieval Workshop (DIR 2012). Ghent University, 2012.

[20] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman and Franciska de Jong. Improving cyberbullying detection with user context. In European Conference on Information Retrieval, pages 693–696. Springer, 2013.

[21] Maral Dadvar, Dolf Trieschnigg and Franciska de Jong. Experts and machines against bullies: A hybrid approach to cyberbullying detection. In Canadian Conference on Artificial Intelligence, pages 275–281. Springer, 2014.

[22] Nektaria Potha and Manolis Maragoudakis. Detecting cyberbullying using time series modeling. In Data Mining Workshop (ICDMW), 2014 IEEE International Conference on, pages 373–382. IEEE, 2014.

[23] Vikas S Chavan and SS Shylaja. A Machine Learning Approach for Detecting Cyberaggressive Peer Comments on a Social Media Network. In Advances in computing, communications and informatics (ICACCI), 2015 International Conference on, pages 2354–2358. IEEE, 2015.

[24] Youssef Bassil and Mohammad Alwani. Error correction algorithm after speech editing