Pakistan Journal of Engineering and Technology, PakJET
Multidisciplinary | Peer Reviewed | Open Access
Volume: SI, Number: 01, Pages: 45- 50, Year: 2020

Cyberbullying Detection Using Machine Learning

Aaminah Ali[1,*], Adeel M. Syed[2]

[1]Software Engineering Department, Bahria University, Islamabad, Pakistan
[2]Software Engineering Department, Bahria University, Islamabad, Pakistan

*Corresponding author: Aaminah Ali (aaminahali@ymail.com).

***Abstract*: It is an age of the Internet and electronic media, and social media platforms are one of the most frequently used communication medium nowadays. But some people use these sites for malicious purpose and among those negative aspects "Cyberbullying" is prevalent. Cyberbullying is a form of bullying done through electronic means and is used to insult or harm others. Many researchers have proposed solutions and strategies to overcome this menace, but sarcasm is one aspect of it that still needs to be touched. This study aims to highlight previous researchers and to propose an approach to detect cyberbullying along with the element of sarcasm included in it. The results proved that SVM classifier performed better than other classifiers.**

***Index Terms*-- Cyberbullying Detection, Machine Learning, Social Media, Text Classification**

I. INTRODUCTION

Social media is a platform which enables users to communicate and interact with their friends online and allows them to share their photos, videos and daily updates. Nowadays, almost every person is found on social media. According to statistics nearly 2 billion users used social sites in 2015 and the figure has now increased to 3.196 billion [1]

Social media has its perks, but it has negative aspects as well. "Cyberbullying" is one of those aspects that need to be handled. Cyberbullying is a form of bullying or harassment that is done through electronic means and is common among young people and teenagers. It includes posting malicious and harmful comments or posts online and sharing personal information about someone to humiliate them[2]. Cyberbullying is one critical issue that is prevailing throughout the world. The person being bullied falls into depression, causes self-harm and in worst cases commits suicide. Thus, Cyberbullying is a severe problem that needs to be taken care of considering the severity of damage it causes to an individual's mental well-being. Apart from taking psychological measures, social media networks should take appropriate steps as well. Though many researchers have proposed and implemented machine learning algorithms, they mostly didn't consider sarcasm in the detection process because they think that it's tough to detect sarcasm from text[3]. This is an important research aspect that should be focused as most of the time; bullies use sarcasm to insult others. Sarcasm is often considered as an indirect form of bullying and is often bitter[4]. Sarcasm is a type of bullying that we often don't take seriously and take it as a joke. Thus, this aspect of bullying needs to be considered and taken seriously.

II. LITERATURE REVIEW

Many researchers have proposed mechanisms to detect cyberbullying [5]. Reviewed machine learning algorithms for cyberbullying detection in Arabic social media networks [6]. Conducted a survey to explore work done on detection and prevention of cyberbullying [7]. Presented a task known as SemEval-2019 which was identifying and categorising offensive language in social media [8]. Conducted a study for detecting cyberbullying and cyber aggression in social media [9]. Audited an existing algorithm using Twitter dataset. The algorithm aims to detect who is the recipient or who is the person being bullied [10]. Proposed a system to monitor cyberbullying by combining message classification and network analysis [11]. Proposed an approach to identify and categorise offensive language on social media [12]. Proposed a multilingual system for detection of cyberbullying as according to them so far courses have been focusing on English language only [13]. Proposed an approach to detect cyberbullying in Instagram media sessions accurately and timely. They named their approach "Concise" [14]. Conducted a review of automated detection techniques for cyberbullying [15]. Proposed a machine learning approach based on SVM classifier utilising a rich feature set [16]. Conducted a research on cyberbullying detection and build their system known as Samurai [3]. Conducted a systematic review of automated approaches of cyberbullying detection [17]. Presented an overview on cyberbullying detection [18]. Proposed a deep learning based on Convolutional Neural Networks for cyberbullying detection [19]. Used deep understanding to detect cyberbullying on different social media platforms [20]. Proposed an approach to detect cyberbullying comprising of three stages. The three stages are aggressive text detection, aggressor and victim detection and cyberbullying case detection. [21] presented a model for cyberbullying detection from social networks. Along with text, they aim to detect cyberbullying from audio, video and image [22]. Conducted a review of cyberbullying detection techniques in social media [23]. Proposed an approach to aggressive and bullying behavior on Twitter [24]. Proposed a multilingual system for cyberbulling detection using machine learning. For this purpose they considered Arabic as well as English language.

III. MATERIALS AND METHODS

This section provides details about the implemented approach. Coming to the implemented approach that is shown in Fig 1. It depicts how cyberbullying was detected in the acquired dataset.

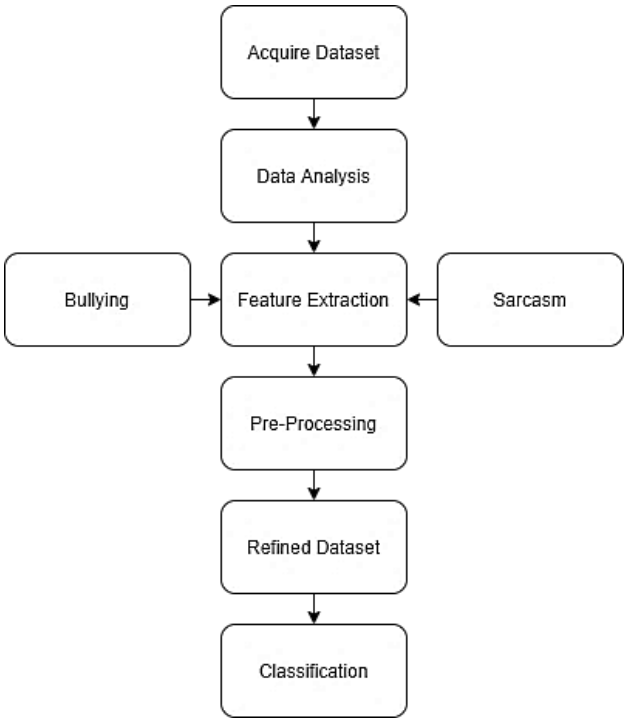

FIGURE 1: Implemented Approach

First o f all data was collected. For this purpose already labelled datasets were acquired that are publicly available on the Internet. These datasets were searched using keywords like cyberbullying, dataset, social media etc. Almost 3-4

datasets were downloaded. These datasets are described in the "Results and Discussion" section.

Before applying preprocessing, some features were extracted that require the data to be in its original form. To detect cyberbullying from the text, features that were extracted are sentiment score, and profanity. For Sarcasm detection count of various features were extracted like "Exclamation Marks", "Question Marks", "Repeated Letters", "Capital Letters", "Intense Adjectives" and "Interjections". These features are considered to be important when detecting sarcasm from the text. [25][26].

After extracting features preprocessing was applied to include the original text in the feature set as well. Preprocessing was performed on all twitter and Formspring datasets. The preprocessing included removal of special characters, single characters left after removing special characters, substituting multiple spaces with single spaces and stopwords. The text was converted into lowercase as well. After that classification was performed.

IV. RESULTS AND DISCUSSION

As machine learning algorithms use numeric data for training, so the text was first converted into the numerical form using a label encoder. After that, the dataset was divided into 80% training set, and 20% test set and then classification algorithms were applied. For classification machine learning algorithms; SVM, naïve Bayes, Random Forest and then an ensemble approach was used. The ensemble approach was a hybrid model consisting of all the algorithms as mentioned above. In this approach, a soft voting criterion was used, which predicts the class label utilising the maximum sum of the predicted probabilities. The results obtained after applying algorithms on datasets are mentioned below:

Table 1 - Dataset 1- Formspring

Algorithm	Accuracy
Random Forest	91%
Naïve Bayes	87%
SVM	92%
Logistic Regression	92%
Ensemble	92%

The table "Dataset 1–Formspring" shows the accuracy achieved after applying classification on the Formspring Dataset. This dataset contains 13,110 posts labelled as bullying and non-bullying. This dataset was created by [27]. These researchers performed various tasks on the datasets like text classification, role labelling, sentiment analysis, as well as topic modelling. They applied machine learning classifiers, i.e. SVM, Naïve Bayes and Logistic Regression to identify bullying traces from the dataset. According to them, SVM performed better than the rest with the accuracy of 81.6%.

Table 2-Dataset 2-Twitter

Algorithm	Accuracy
Random Forest	70%
Naïve Bayes	71%
SVM	73%
Logistic Regression	73%
Ensemble	73%

The table "Dataset 2–Twitter" depicts the accuracy achieved after applying classification on the Twitter dataset consisting of 13,420 tweets labelled as "offensive" and "not offensive" by [28]. These authors also applied three machine learning classifier on this dataset, i.e. SVM, Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Network (CNN). They used precision, recall and F1-score as metrics. According to their results, CNN performed the best in classifying tweets as offensive or not offensive.

Table 3- Dataset 3- Twitter

Algorithms	Accuracy
Random Forest	68%
Naïve Bayes	69%
SVM	72%
Logistic Regression	72%
Ensemble	72%

The table "Dataset 3–Twitter" depicts the accuracy achieved after applying classification on the Twitter dataset. It includes 8817 tweets labelled as either positive (bullying) or negative (non-bullying)

Table 4-Dataset 4-Twitter

Algorithm	Accuracy
Random Forest	78%
Naïve Bayes	76%
SVM	80%
Logistic Regression	78.6%
Ensemble	79.5%

The table "Dataset 4–Twitter" depicts the accuracy achieved after applying classification on a Twitter hate-offensive dataset. It contains 24,784 tweets labelled as offensive, hate or none. This dataset was used by [29], and according to the SVM and Logistic Regression performed better than other models, i.e. decision trees, random forest and Naïve Bayes. Our results also depict the same as SVM performed better than others.

From the tables, it is evident that in almost all datasets the accuracy achieved by SVM, Logistic and Ensemble approach is the same except for Dataset 4 where SVM classifier took the lead. After completing accuracy for each dataset separately, the average was calculated. For example, the average accuracy for Random Forest was calculated by adding accuracy achieved on all datasets divided by the total number of datasets, i.e. 4. The average accuracy was calculated to analyse which classifier performs the best overall. The table below depicts the average accuracy of classifiers against all datasets.

Table 5--Average Accuracy

Classifier	Average Accuracy
Random Forest	76.7 %
Naïve Bayes	75.7%
SVM	79.3%
Logistic Regression	78.9%
Ensemble	79 %

The results clearly show that the used approach yielded considerably good results and by observing the average accuracy we can say that SVM and Ensemble classifier performed better than others.

V. CONCLUSION

This particular study aimed to explore cyberbullying detection using machine learning. The previous work done in this regard was also highlighted. Cyberbullying is a vast term and has different aspects. Among those aspects, sarcasm is essential. Sarcasm is a way of insulting someone and has adverse effects on the victim. As per our observation, this aspect of bullying was not considered in the previous researches. Hence this study aimed to include that aspect as well. In this study, we detected cyberbullying using machine learning algorithms. Then results were presented in a tabular format. The results depicted that SVM and Ensemble performed better than remaining classifiers with 79% average accuracy. The second one is Logistic Regression with 78% accuracy, then Random Forest with 76.7% while Naïve Bayes performed classification with 76% accuracy. However, in this study, only textual features were considered. For future, network and contextual features can be considered.

VI. ACKNOWLEDGEMENT

First and foremost, thanks to Allah Almighty, the Most Beneficent and the Most Merciful, for giving me the ability and opportunity to learn and for His infinite blessings. Then I would like to thank my parents for their utmost support and belief in me.

REFERENCES

[1] Chaffey Dave, "Global social media research summary 2019 | Smart Insights," 2019. [Online]. Available: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/. [Accessed: 28-Sep-2019].

[2] "What Is Cyberbullying | StopBullying.gov." [Online]. Available: https://www.stopbullying.gov/cyberbullying/what-is-it/index.html. [Accessed: 30-Sep-2019].

[3] and J. L. Semiu Salawu, Yulan He, "Approaches to Automated Detection of Cyberbullying: A Survey," in *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, 2017, pp. 3–24.

[4] "Think Sarcasm is Funny? Think Again | Psychology Today." [Online]. Available: https://www.psychologytoday.com/us/blog/think-well/201206/think-sarcasm-is-funny-think-again. [Accessed: 05-Dec-2019].

[5] B. Y. Alharbi, M. S. Alharbi, N. J. Alzahrani, M. M. Alsheail, and D. M. Ibrahim, "Using Machine Learning Algorithms for Automatic Cyber Bullying Detection in Arabic Social Media Using Machine Learning Algorithms for Automatic Cyber Bullying Detection…," *J. Inf. Technol. Manag.*, vol. 12, no. 2, pp. 123–130, 2020.

[6] V. Krithika and V. Priya, "A Detailed Survey On Cyberbullying in Social Networks," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020, pp. 1–10.

[7] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Identifying and Categorising Offensive Language in Social Media (OffensEval)," in *Proceedings ofthe 13th International Workshop on Semantic Evaluation (SemEval-2019)*, 2019, pp. 75–86.

[8] D. Chatzakou *et al.*, "Detecting Cyberbullying and Cyberaggression in Social Media," *ACM Trans. Web*, vol. 13, no. 3, Jul. 2019.

[9] V. K. Singh and C. Hofenbitzer, "Fairness across network positions in cyberbullying detection algorithms," 2019.

[10] S. Menini *et al.*, "A System to Monitor Cyberbullying based on Message Classification and Social Network Analysis," in *Proceedings ofthe Third Workshop on Abusive Language Online*, 2019, pp. 105–110.

[11] S. D. Swamy, A. Jamatia, B. Gambäck, and A. Das, "An Ensemble Approach to Identifying and Categorising Offensive Language in Twitter Social Media Corpora," in *Proceedings ofthe 13th International Workshop on Semantic Evaluation (SemEval-2019)*, 2019, pp. 696–703.

[12] P. Rohit and R. R. Rajeev, "Multilingual Cyberbullying Detection System," in *IEEE International Conference on Electro Information Technology (EIT)*, 2019, pp. 40–44.

[13] M. Yao, C. Chelmis, and D. S. Zois, "Cyberbullying ends here: Towards robust detection of cyberbullying in social media," in *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 2019, pp. 3427–3433.

[14] C. T. and P. O. Thabo Mahlangu and Institute of Electrical and Electronics Engineers, "A Review of Automated Detection Methods for Cyberbullying," in *International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, 2018, p. 5.

[15] C. Van Hee *et al.*, "Automatic detection of cyberbullying in social media text," *PLoS One*, vol. 13, no. 10, Oct. 2018.

[16] M. Ptaszyński, G. Leliwa, M. Piech, and A. Smywiński-Pohl, "Cyberbullying Detection -- Technical Report 2/2018, Department of Computer Science AGH, University of Science and Technology," Aug. 2018.

[17] Wan Noor Hamiza Wan Ali, Masnizah Mohd, and Fariza Fauzi, "Cyberbullying Detection: An Overview," in *Proceedings of the 2018 Cyber Resilience Conference (CRC)* , 2018, pp. 13-15,.

[18] M. Abdullah Al-Ajlan and M. Ykhlef, "Deep Learning Algorithm for Cyberbullying Detection," *IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 9, 2018.

[19] S. Agrawal and A. Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms," in *ECIR*, 2018.

[20] L. P. Del Bosque and S. E. Garza, "Cyberbullying Detection in Social Networks: A Multi-Stage Approach," *Res. Comput. Sci.*, vol. 148, no. 3, pp. 285–296, 2018.

[21] PradheepT PG, JISheeba, Pradeep Devaneyan, and Yogeshwaran.T, "AUTOMATIC MULTIMODEL CYBERBULLYING DETECTION FROM SOCIAL NETWORKS," in *Proceedings of the International Conference on Intelligent Computing Systems* , 2017, pp. 1556–5068.

[22] Sonika Sharavastav, "A Review of Cyberbullying Detection in Social Networking," in *Information,Communication and Computing Technology(ICICCT)*, 2017.

[23] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean Birds: Detecting Aggression and Bullying on Twiier," in *ACM*, 2017, pp. 13–22.

[24] B. Haidar, M. Chamoun, and A. Serhrouchni, "A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning," vol. 2, no. 6, pp. 275–284, 2017.

[25] Pushpak Bhattacharyya, "Sarcasm Detection: A Computational and Cognitive Study," California, 2018.

[26] K. Sundararajan and A. Palanisamy, "Multi-Rule Based Ensemble Feature Selection Model for Sarcasm Type Detection in Twitter," *Comput. Intell. Neurosci.*, vol. 2020, p. 17, 2020.

[27] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from Bullying Traces in Social Media," in *Proceedings of the 2012*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2012, pp. 656–666.

[28] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the Type and Target of Offensive Posts in Social Media," in *Proceedings of NAACL-HLT*, 2019, pp. 1415–1420.

[29] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, 2017, no. Icwsm, pp. 512–515.