Detecting Cyber Defamation in Social Network Using Machine Learning

Kowsalya Devi.V
PG Student, Dr. N.G.P Arts and Science College

Abstract **- Cyber Defamation is the process of sending wrong messages to a person or community which causes heated debate with users. Cyber defamation is mostly seen in social networking sites where users reply to post with bullying words to threaten or insult other users. Cyber defamation is considered a misuse of technology. According to the newest survey done on all over the world data day by day, cases are enlarging on cyberbullying. In order to solve this problem many natural language processing techniques are preferred by various authors which are time taking and not automatic. With the advancement of machine learning and artificial intelligence, models can be manufacture and automatic detection can be executed. To show this scenario live social media application is developed in python programming with and the Naive Bayes algorithm is used to train the model on a Social Media dataset and using this model live detection of cyberbullying is predicted and alert messages are shown on the application.**

Index Terms **- Detecting Cyber Defamation, Machine Learning, Naive Bayes algorithm.**

I.INTRODUCTION

Social networking sites are great tools for joining with people. However, as social networking has become outspread, people are finding illegal and corrupt ways to use these communities. We see that people, especially teens and youngster, are finding new ways to bully one another over the Internet. Close to 25% of parents in a study guided by Symantec revealed that, to their knowledge, their child has been involved in a cyberbullying incident. There are no well-known datasets for research on cyberbullying. A set of huge datasets was made available at the Content Analysis on the Web workshop for a misdeed detection task; however, this dataset was unlabeled. Furthermore, the data was pulled from a collection of sources and with the exception of the data from Kongregate. The datasets appear to be discussions among adults. In order to conduct the study reported herein, we developed our own labeled dataset containing data from a web crawl of Forum spring. Social networking has end up a well-known recreation within the web at present, attracting hundreds of thousands of users, spending billions of minutes on such services When making use of Social network's (SN's), one of a kind men and women share one-of-a-kind quantities of their private understanding During the earlier years, online identification theft has been a primary problem considering it affected millions of people's worldwide. Victims of identification crime may suffer unique types of damages for illustration, they would drop time or cash, get dispatched to reformatory, get their public image ruined, or have their relationships with associates and loved ones damaged. False profiles are the profiles which are not specific. They are the profiles of men and women with false credentials. The false Facebook profiles more commonly are allowing in malicious and unattractive activities, causing problems to the social community customers. Individuals create fake profiles for social engineering, online enactment to defame a man or woman, promoting and agitate for a character or a crowd of individuals. Facebook has its own security system to secure person authorization from spamming, phishing, and so on. The problems involving like social networking like privacy, online bullying, misuse, and trolling and many others. With the rise of so called trends of sharing or posting dates or pictures on certain social networking sites and commenting on them have adding the risk of cyber defamation. The generally used social media brought a revolt not only in the Indian sphere but also all over the world. The remarkable growth of the internet has provided people with the platforms to express.

II. METHODOLOGY

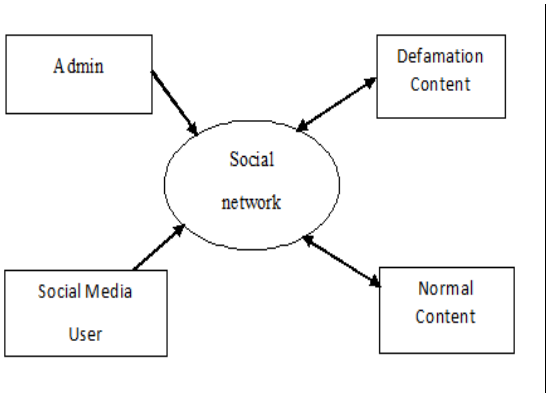

Fig 1: Block diagram for Social Network

The present proposed system modules are developed using Python and MYSQL. Detecting Defamation over Social Media is an area for like-minded people to exchange ideas, post articles, comment articles, offer likes on relevant posts and share their ideas with others also. This system also provides ways of archiving (or storing) and searching for previous exchanges. It is an online service that allows registered users to post articles and responses to others timeline. This project also deals with the sending/receiving comments between the registered users. To reduce the drawback in the existing system, new software is developed in a user-friendly manner to satisfy and overcome the drawback. The defamation technique monitors every single post happens in this social media and every share will be administered by the automated system. Natural language processing technique combined with keyword matching algorithm ensures identifying defamed profiles and odd out and list them to admin user. This project we will develop using python and machine learning. Within that first we will search and find the dataset and download it for train the model. After downloading first, we will pre-treat the data and then fetch to Tf-Idf. Then with the help of naïve bayes, we train the dataset and generate model separately. Then we are going to develop a web-based application using FLASK framework. We will fetch the real time tweets from twitter and then we apply generated model to these fetched tweets and check the text or images are cyberbullying or not. These all-purpose we are using python as backend, MySQL is database and for frontend html, CSS, JavaScript.

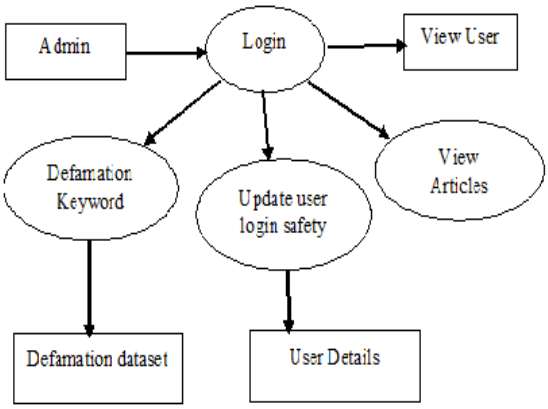

Fig 2:Block diagram of Login and User

III. LITERATURE REVIEW

There are many attitudes that proposes systems which can detect cyberbullying automatically with high exact. First one is author Nandhini et al [1]. have proposed a model that uses Naïve Bayes machine learning approach and by their work they achieved 91% accuracy and got their dataset from MySpace.com, and then they proposed another model [2]. Naive Bayes classifier and genetic operations (FuzGen) and they achieved 87% accuracy. Another approach by Romsaiyudet al [3]. they enhanced the Naïve Bayes classifier for extracting the words and examining loaded pattern clustering and by this approach they achieved 95.79% accuracy on datasets from Slashdot, Kongregate, and MySpace. However, they have a problem that the cluster processes do not work in parallel. Moreover, in the accurate proposed by Bunchanan et al[4].they used War of Tanks game chat to get their dataset and manually classified them and then differentiate them to simple Naïve classification that uses sentiment analysis as a feature, their results were poor when compared to the manually classified Results. Furthermore, Isa et al[5]. proposed an approach after getting their dataset from kaggle they used one classifier Naïve Bayes. The Naïve Bayes classifier yielded average accuracy of 92.81% while SVM with poly kernel yielded accuracy of 97.11%, but they did not mention their training or testing size of the dataset, so the results may not be acceptable. Another Approach by Dinakar et al[6]. that aimed to detect explicit bullying language pertaining to Sexuality, Race & Culture and intelligence, they acquired their dataset from YouTube comment

section. After applying Naïve Bayes classifiers, SVM yielded accuracy of 66% and Naïve Bayes 63%. Moving on to Di Capua et al[7]. they proposed a new way for cyberbullying detection by adopting an unsupervised approach, they used the classifiers inconsistently over their dataset, applying SVM on Form Spring and achieving 67% on recall, applying GHSOM on YouTube and achieving 60% precision, 69% accuracy and 94% recall, applying Naïve Bayes on Twitter and achieving 67% accuracy. Additionally, Haidar et al[8]. proposed a model to detect cyberbullying but using Arabic language they used Naïve Bayes and achieved 90.85% precision and SVM achieved 94.1% as precision but they have extraordinary of false positive also they are work on Arabic language. Another type of proposal using machine learning. One of the proposed methods is Zhang et al. [9] in their paper uses novel pronunciation-based convolution neural network (PCNN), thereby lighten the problem of noise and bullying data scarcity to overcome class imbalance. 1313 messages from twitter, 13,000 messages from spring. Accuracy of the twitter dataset was not calculated thanks to it being imbalanced. While Achieving 56% on precision, 78% recall and 96% accuracy, while achieving high accuracy their dataset was unbalanced, so that gives false results and that reflects in precision score which is 56%. The authors Nobata et al. [10] showed that using abusive language has enlarged recently, they used a framework called Vowpal wabbit for classification, and they also developed a supervised classification methodology with NLP features that execute machine learning approach, The F-Score reached 0.817 using dataset collected from comments posted on Yahoo News and Finance. Zhao et al. [11] proposed framework particular for cyberbullying detection, they used word embedding that creates an inventory of pre-defined insulting words and allocate weights to get bullying features, they used SVM as their main classifier and got recall of 79.4%. Then another approach was proposed by Parime et al. [12] they got their dataset from MySpace and manually marked them, and they used the SVM Classifier for the classification. Moreover, Chen et al. [13] proposed a replacement feature extraction method called Lexical Syntactic Feature and SVM as their classifier and that they reached 77.9% precision and 77.8% recall. Furthermore, Ting et al. [14] proposed a strategy based on SNM, they collected their data from social media and then used SNA calculation and sentiments as features. Seven experiments were made, and they reached around 97% precision and 71% as recall. Furthermore, Harsh Dani et al. [15] introduced a new framework called SICD, they used KNN for classification. Finally, they achieved 0.6105 F1 score and 0.7539 AUC score. SVM classifier was one of the proposals used in the research papers.

IV. CONCLUSION

In this paper, we proposed an approach to identify the fake profile in social network using limited profile data, about 2816 users. As we concluded in our paper, we demonstrate that with limited profile data our approach can identify the fake profile with 99.64% correctly classified instances and only 0.35% incorrectly classified instances, which is comparable to the results obtained by other existing approaches based on the larger data set and more profile information. Our research can be a motivation to work on limited social network information and find solutions to make better decision through authentic data. Additionally, we can attempt similar approaches in other domains to find successful solutions to the problem where the least amount of information is available. In future work we expect to run our model using more sophisticated concepts such as ontology engineering, in order to semantically analyze user posts, and comportments. This later concept can improve the quality of prediction of fake or genuine profiles.

REFERENCES

[1] B Nandhini and JI Sheeba. Cyberbullying detection and classification using information retrieval algorithm. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), page 20. ACM, 2015.

[2] B Sri Nandhini and JI Sheeba. Online social network bullying detection using intelligence techniques. Procedia Computer Science, 45:485–492,2015.

[3] Walisa Romsaiyud, Kodchakornna Nakornphanom, Pimpaka Prasertsilp, Piyaporn Nurarak, and Pirom Konglerd. Automated cyberbullying detection using clustering

appearance patterns. In Knowledge and Smart Technology (KST), 2017 9th International Conference on, pages 242–247. IEEE, 2017.

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

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

[6] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3):18, 2012.

[7] Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino. Unsupervised cyber bullying detection in social networks. In Pattern Recognition (ICPR), 2016 23rd International Conference on, pages 432–437. IEEE, 2016.

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

[9] Xiang Zhang, Jonathan Tong, Nishant Vishwamitra, Elizabeth Whittaker, Joseph P Mazer, Robin Kowalski, Hongxin Hu, Feng Luo, Jamie Macbeth, and Edward Dillon. Cyberbullying detection with a pronunciation based convolutional neural network. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 740–745. IEEE, 2016.

[10] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In Proceedings of the 25th international conference on world wide web, pages 145–153. International World Wide Web Conferences Steering Committee, 2016.

[11] Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on bullying features. In Proceedings of the 17th international conference on distributed computing and networking, page 43. ACM, 2016.

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

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

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

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