International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 794-798 www.ijaem.net ISSN: 2395-5252

Social Media Content Filtering using Machine Learning Techniques

Prof.RahulM. Raut[1], Shubham Kandekar[1] Parag Thorat[2]
Gayatri Jawharkar[3], Sarvesh Sonawane[4],
[1,2,3,4,5]*Information Technology Department, Sandip Institute Of Technology And Research Centre Nashik ,India*

Date of Submission: 15-02-2023 Date of Acceptance: 25-02-2023

ABSTRACT:This research paper is about the important problem of rising hate and offensive Contents made against people or communities on social media is addressed by this proposed system. Such behaviour has become widespread in social media, where people can easily express their hatred and connect with many people they might not otherwise consider in the real world. The use of computational techniques to identify such offensive and hateful content and to take action against it is one of the most efficient ways to address this enigmatic problem. Since English is the most widely used language on the Internet, the current work focuses on identifying hate speech and offensive material in english and other regional languages.To the best of our knowledge, no study has been conducted to compare the various feature engineering techniques and machine learning algorithms in order to determine which feature engineering technique and machine learning algorithm outperform on a standard publicly available dataset. As a result, in this work, we compare the performance of various feature engineering techniques and machine learning algorithms on a publicly available dataset before developing a cloud-based application.

Keywords: Hate speech; online social networks; natural language processing; text classification; machine learning

I. INTRODUCTION

In this paper the Project App refers to irrelevant or unsolicited messages sent over the messengers for abusing or may harm someone's personal life. The spam may or may not be harmful to the intended people. Message Protection is a tedious task and in the absence of automatic measure for filtering of message, the task of spam filtering is taken up with the person at the receiving end. We will classify the spam comment of the social media platforms. We use the machine learning concept which is a subset of artificial intelligence. Four types of machine learning modules are available, namely supervised learning, semi-supervised learning, unsupervised learning, and strengthening. Machine learning is the method of extraction, transforming, loading and predicting the meaningful information from huge data to extract some patterns and also transform it into understandable structure for further use. Prediction and classification and are the two kinds of data analysis techniques that are used by mine models that identify the most relevant data classes and predict future data trends.

II. RELATED WORK

Due to the popularity of social media, such as Twitter and Sina Weibo, many research works have been conducted on spam comments detection. The existing research mainly focused on two aspects: detecting spammers and analyzing content features of spam comments. Some researchers detected spammers by analyzing social networking. Stringhini et al. considered that spammers have unreasonable social networking relations compared with normal users, such as following lots of users but having less fans. A. Author: Gianluca et al. Stringhini. Detecting spammers on social networks Proceedings of the 26th Annual Computer Security Applications Conference, ACM, 2010. In this paper, we analyze to which extent spam has entered social networks. More precisely, we analyze how spammers who target social networking sites operate. To collect the data about spamming activity, we created a large and diverse set of "honey-profiles" on three large social networking sites, and logged the kind of contacts and messages that they received. We then analyzed the collected data and identified anomalous behavior of users who contacted our profiles. Based on the analysis of this behavior, we developed techniques to detect spammers in social networks, and we aggregated their messages in large spam campaigns. Our results show that it is

possible to automatically identify the accounts used by spammers, and our analysis was used for take-down efforts in a realworld social network. More precisely, during this study, we collaborated with Twitter and correctly detected and deleted 15,857 spam profiles. Some researchers detected spammers by analyzing social networking. Stringhini et al. considered that spammers have unreasonable social networking relations compared with normal users, such as following lots of users but having less fans [1]

A. Author: Chengfeng Lin et al. Analysis and identification of spamming behaviors in sinaweibo microblog. In Proceedings of the 7th Workshop on Social Network Mining and Analysis. ACM, 2013. Jong Myoung Kim, Zae Myung Kim, and KwangjoKim. An approach to spam comment detection through domain-independent features. In Big Data and Smart Computing (BigComp), 2016 International Conference on,. IEEE, 2016. Several recent studies have focused on detecting spammers on Sina Weibo For example, first identifies three representative spamming behaviors: aggressive advertising, repeated duplicate reposting and aggressive following, and subsequently develops automated spamming behavior classifiers to filter spammers in Sina Weibo, while implements 50 honeyspots on two most popular microblogging services in China: Sina Weibo and Tencent Weibo, and examines spammers'sbehavior from a variety of perspectives including social information, activity, account age and spamming strategies. Rather than inferring Weibo spammers, presents a machine-learning based algorithm for identifying individual spam messages However, it's not difficult for spammers to get lots of ossified fans. Thus, these methods based on social networking may not work well. Some researchers detected spammers by analyzing the users' attributes and representative behaviors, such as registration date, repeated reposting and aggressive following [2,3]

C. Author: Chenwei Liu, Jiawei Wang, and Kai Lei. Detecting spam comments posted in micro-blogs using the self-extensible spam dictionary. In 2016 IEEE International Conference on Communications (ICC), pages 1–7. IEEE, 1 st May 2016. Our experimental results demonstrate that when detecting a combination of both AD and vulgar spam comments, we can achieve an average detection accuracy of 87.9%. Particularly for AD spam comments detection, we can achieve an average accuracy of 96.2%, which is preferable compared to when using machine learning methods. The statistical analysis of the results verifies that our proposed methods can identify the spam comments effectively and to relatively high degrees of accuracy. However, from our observation, lots of normal comments are short and inconsistent with microblogs' theme. Liu et al. proposed spam dictionary and Proportion-Weight Filter model to detect two kinds of spam comments (advertisement and vulgar comments), and achieved an average accuracy value of 87.6%[4]

D. Author: Archana Bhattarai, Vasile Rus, and Dipankar Dasgupta. Characterizing comment spam in the blogosphere through content analysis. In Computational Intelligence in Cyber Security, 2009., IEEE, 2009. Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. Survey of review spam detection using machine learning techniques. In Journal of Big Data, volume 23, 2015. They focused their analysis purely on comments and did not correlate the comments with their corresponding blog posts. In 2009, Bhattarai et al performed content analysis of spam comments to identify features such as number of word duplications, stop words ratio etc., which were used to train classifiers for spam detection. They obtained an accuracy of 86% in detecting spam comments using their approach Results have much room for improvement because they ignored other text-based features. Romero et al. performed a comparative study using four classifiers (Naive Bayes, K-Nearest Neighbors, Neural Networks and Support Vector Machines) in spam comments detection. Support Vector Machines got the highest performance of 84.6%[5,6]. By extracting several useful features, these machine learning methods get decent results. However, with lots of new cyber words created daily, these methods may not understand the ever-changing expression.

E. Author: Fangzhao Wu, Jinyun Shu, Yongfeng Huang, and Zhigang Yuan. Co-detecting social spammers and spam messages in microblogging via exploiting social contexts. In Neurocomputing, volume 201, 2016. In this paper, we propose a unified approach for social spammer and spam message co-detection in microblogging. Our approach utilizes the posting relations between users and messages to combine social spammer detection with spam message detection. In addition, we extract the social relations between users and the connections between messages to refine detection results. We regard these social contexts as the graph structure over the detection results and incorporate them into our approach as regularization terms. Besides, we introduce an

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 794-798 www.ijaem.net ISSN: 2395-5252

efficient optimization algorithm to solve the model of our approach and propose an accelerated method to tackle the most time-consuming step. Extensive experiments on a real-world microblog dataset demonstrate that our approach can improve the performance of both social spammer detection and spam message detection effectively and efficiently.[7]

F. Author: Mahmoud, T.M. and Mahfouz, A.M. (2012) SMS Spam Filtering Technique Based on Artificial Immune System. IJCSI International Journal of Computer Science Issues, 9, 589-597. In this paper, analysis of different types of phishing attacks on mobile devices is provided. Mitigation techniques—anti-phishing techniques— are also analyzed. Assessment of each technique and a summary of its advantages and disadvantages is provided. At the end, important steps to guard against phishing attacks are provided. The aim of the work is to put phishing attacks on mobile systems in light, and to make people aware of these attacks and how to avoid them.[8]

III. PROBLEM STATEMENT

• Spam comments refer to the unwanted comments with rude words, advertisement, political or religious views. Massive spam comments seriously decrease users' reading experience and hinder the healthy development of social media. Thus, it is essential to detect and filter spam comments.

• Due to increase in usage of social media networks crime rate has been increased.

• Such as Abusive words, misinformation, some fake URLs.

• Even spreading of offensive content like pornography is been increased.

• Due to which unhealthy weather is been created on social media by which one can easily get depressed.

IV. OBJECTIVES

The objectives of the system are

• To develop an automated deep learning-based approach for detecting hate speech and offensive language

• To Create a highly accurate machine learning model to classify toxic comments and images on the social media and messaging platforms

• To reduce malicious activities such as harassing a person, identity theft, and privacy violations using machine learning techniques.

• To deploy the application on the cloud.

• To test and validate the results.

V. PROPOSED WORK

The detection of hate speech is a difficult problem. There are virtually limitless ways for people to express themselves, including hate speech. As a result, composing rules or a list of hate words by hand is impossible, so we created a system based on machine learning algorithms. The main goal of the project is to create an application that can process social media messages and block the most likely hate speech messages for manual inspection. As a result, we needed to devise a method for identifying potential hate speech messages and training the hate speechdetector during the experiment period. Our method consists of several processes. The first process is preprocessing then feature extraction followed with classification and model evaluation.

• Dataset Hate speech tweets are used as a dataset in this study.The information provided in the dataset includes the label whether the tweet is abusive tweet or not abusive tweet, whether it is hate speech or not hate speech, the hate speech categories, target, and level. Hate speech always aims at a specific target. Generally, the target of hate speech consists of two kinds of target. Those targets are toxic speech and Non-toxic. The categories information for the dataset is based on the topic of the tweet such as gender, religion, physical, race, and others. The level of hate speech is divided into weak, moderate, and strong hate speech. For this study, we only used the dataset which has a target label. The dataset consists of more tha 5000 hate speech tweets divided into two class label HATE and NON HATE. As many as 3000 tweets are labeled as HATE and 2000 tweets are labeled as NON HATE. The dataset then divided into two, 80% for training data and 20% for testing data. The splitting process is done after the dataset preprocessing. Then the features from the dataset are extracted and use in the classification process.

• Preprocessing The first process in hate speech target classification is to preprocess the dataset. This process was done to make the dataset clean of noises and prepared to be used in the classification process. Several preprocessing steps were used in this study. Those steps consist of emoji removal, case folding, special character removal, stop-word removal, and stemming.

• Feature Extraction After the preprocessing process was done, features from the dataset were extracted into a feature vector. We used several word n-gram features such as unigram, bigram, trigram, and combination of unigram, bigram, and trigram. Two-term weighting schemes were used for the feature extraction

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 2 Feb. 2023, pp: 794-798 www.ijaem.net ISSN: 2395-5252

process. The term weighting schemes used were Bag-of-Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF).

• Classification We implemented several machine learning algorithms as the classifier for taret classification of hate speech in tweets. Those algorithms are Support Vector Machine (SVM) and Naive Bayes (NB) According to the previous study of hate speech classification, The training phase used 80% of the dataset as the training data, while the testing phase used the remaining 20% of the dataset as the testing data and CNN for the images classification.

• Evaluation The evaluation measurement used in this study is F1- score. Accuracy is not used as evaluation measurement because it cannot guarantee that high accuracy shows that the model can predict well considering the accuracy paradox. F1-score is obtained by calculating harmonic mean between precision and recall.

VI. CONCLUSION

In this work, we present the target classification of hate contents for social using machine learning. We are using two algorithms as a classifier to compare the results. Those algorithms are Na¨ıve Bayes/SVM for text and CNN for images Two-term weighting schemes including Bag-of-Words and TF-IDF are used and several word n-grams are used as feature representations. From previous work and literature survey we found that using Bag-of-Words as term weighting scheme, the best result achieved F1-score more than 0.85 using SVM with word unigram as the feature. When using TF-IDF as term weighting scheme, the best result achieved F1-score more than 0.82 also using SVM and word unigram as the feature. The results show that word unigram is good feature representation for hate speech target classification in Twitter. In terms of term weighting scheme, TF-IDF achieved a slightly better result than Bag-of-Words with a difference in F1-score of 0.00249. According to the results obtained from previous work, we can conclude that SVM is the best algorithm for hate speech target classification. hence we are using SVM and Naive bays algorithm for this work.

VII. FUTURE SCOPE

For future study, we suggest implementing a deep learning approach to obtain better performance in hate speech target classification. Deep learning is known to perform better than machine learning in text classification. However, in order to use deep learning efficiently, we suggest increasing the number of hate speech target dataset. Using the current dataset might reduce the effectiveness of deep learningbecause of the small number of the dataset. Hopefully, by using deep learning for hate speech target classification, the result will achieve better performance.

REFERENCES

[1]. Gianluca et al. Stringhini. Detecting spammers on social networks. In Proceedings of the 26th Annual Computer Security Applications Conference, pages 1–9. ACM, 2010.

[2]. Chengfeng Lin et al. Analysis and identification of spamming behaviors in sinaweibo microblog. In Proceedings of the 7th Workshop on Social Network Mining and Analysis. ACM, 2013.

[3]. Jong Myoung Kim, Zae Myung Kim, and Kwangjo Kim. An approach to spam comment detection through domain-independent features. In Big Data and Smart Computing (BigComp), 2016 International Conference on, pages 273–276. IEEE, 2016.

[4]. Chenwei Liu, Jiawei Wang, and Kai Lei. Detecting spam comments posted in micro-blogs using the self-extensible spam dictionary. In 2016 IEEE International Conference on Communications (ICC), pages 1–7. IEEE, 2016.

[5]. Archana Bhattarai, Vasile Rus, and Dipankar Dasgupta. Characterizing comment spam in the blogosphere through content analysis. In Computational Intelligence in Cyber Security, 2009., pages 37–44. IEEE, 2009.

[6]. Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. Survey of review spam detection using machine learning techniques. In Journal of Big Data, volume 23, 2015.

[7]. Fangzhao Wu, Jinyun Shu, Yongfeng Huang, and Zhigang Yuan. Co-detecting social spammers and spam messages in microblogging via exploiting social contexts. In Neurocomputing, volume 201, pages 51–65, 2016.

[8]. T. M. Mahmoud and A. M. Mahfouz, "Sms spam filtering technique based on artificial immune system," IJCSI International Journal of Computer Science Issues, vol. 9, no. 1, pp. 589–597, 2012.

[9]. X. Huang and M. Xu, "An Inter and Intra Transformer for Hate Speech Detection," 2021 3rd International Academic

Exchange Conference on Science and Technology Innovation (IAECST), 2021, pp. 346-349, doi: 10.1109/IAECST54258.2021.9695652.

[10]. D. Sahnan, S. Dahiya, V. Goel, A. Bandhakavi and T. Chakraborty, "Better Prevent than React: Deep Stratified Learning to Predict Hate Intensity of Twitter Reply Chains," 2021 IEEE International Conference on Data Mining (ICDM), 2021, pp. 549-558, doi: 10.1109/ICDM51629.2021.00066.

[11]. H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," in IEEE Access, vol. 6, pp. 13825-13835, 2018, doi: 10.1109/ACCESS.2018.2806394.

[12]. S. Alsafari and S. Sadaoui, "Semi-Supervised Self-Learning for Arabic Hate Speech Detection," 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2021, pp. 863-868, doi: 10.1109/SMC52423.2021.9659134.

[13]. K. -Y. Lin, R. K. -W. Lee, W. Gao and W. -C. Peng, "Early Prediction of Hate Speech Propagation," 2021 International Conference on Data Mining Workshops (ICDMW), 2021, pp. 967-974, doi: 10.1109/ICDMW53433.2021.00126.

[14]. R. A. Ilma, S. Hadi and A. Helen, "Twitter's Hate Speech Multi-label Classification Using Bidirectional Long Short-term Memory (BiLSTM) Method," 2021 International Conference on Artificial Intelligence and Big Data Analytics, 2021, pp. 93-99, doi: 10.1109/ICAIBDA53487.2021.9689767.