International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

Profanity Statistical Analyzer

Akshay Limje[1], Ashutosh Patil[2], Nilay Tagde[3], Purvesh Kondewar[4], Prof. Ashish Golghate[5]

[1,2,3,4,5]*Department of Computer Science and Engineering, DBACER, Wanadongri, Hingna Road-441110, Nagpur.*

Abstract - *Nowadays use of internet has become a big trend among the children especially teenagers. Use of internet and various social networking sites has become a daily part of the children's activity. While interacting with the social media websites, they come across many profane/abusive words which are not appropriate for them. Some social media claim that they provide a decent platform for the children, these contents of social media including webpages and blogs may create a bad impact for the people where they might end up in sudden change in attitude which might lead to depression. Therefore, to deal with this problem, we have developed a tool* ***"PROFANITY STATISTICAL ANALYZER"****. This analyzer will help us find the profane/abusive contents from their webpage, social media posts and blogs over the internet. There is a margin for profane words through which we can classify the website, post or blog as inappropriate to the person of particular age. Calculating the percentage amount of profane word will help us to categorize those websites and show the actual profane/offensive words in the contents, at the same time helping the user to decide whether to visit or not to visit the specific webpage.*

***Key Words*: Social Media, Blog, Analyzer, Profane.**

1. INTRODUCTION

Anytime one engages online, whether on message board forums or comments on social media, there is always a serious risk that he or she may be the target of ridicule and even harassment. Nevertheless, major social media companies like Facebook, Twitter and Reddit find it difficult to tackle this problem as the number of posts that contains the abusive contents cannot be eliminated with only human resources. To combat abusive language, many IT companies have standard guidelines that must be followed by the users. They employ human moderators in conjunction with systems that use regular expressions and blacklist, to recognize bad language and thus a post can be removed.

As large number of people communicate online, the need for high quality automated abusive language detectors become much more profound. So, to take necessary actions we have made a solution that will surely help to eradicate this problem. In this project, we are aiming towards an approach of detecting abusive language in the given post, written blogs and web pages so that user must be able to detect the offensive content of all types that are posted on that particular webpage.

Basically, this project is mainly divided into three different phases. In the first phase, we are taking a webpage or a blog with some abusive content as an input. This input will then be tokenized to form a dataset. In second phase, this dataset which is then compared with the dataset containing a list of abusive words with a custom written function in Python that results in finding abusive words. In the third phase, result of the comparison between input dataset and abusive words dataset is used to provide users with essential information that will help them to decide whether to visit or not to that webpage and result may be presented in terms of percentage or in the form of graph or pie charts.

2. LITERATURE REVIEW

1) Y. Chen, Y. Zhou, S. Zhu, and H. 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), IEEE, 2012. :- Their work was one of the first to use a combination of lexical syntactic feature-based language model and parser features to detect offensive language in YouTube comments to shield adolescents. While they do note that they do not have a strict definition of offensive language in mind, their tool can be tuned by the use of a threshold which can be set by parents or teachers so online material can be filtered out before it appears on a web browser. The work takes a supervised classification approach using Support Vector Machines (SVMs) with features including n-grams, automatically derived blacklists, manually developed regular expressions and dependency parse features. They achieve a performance on the task of inflammatory sentence detection of precision of 98.24% and recall of 94.34%.

2) Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017), "Deep learning for hate speech detection in tweets." Proceedings of the 26th International Conference on World Wide Web Companion, April-3-2017: -

They have selected few baseline methods and then discuss the proposed approach. In all these methods, an embedding is generated for a tweet and is used as its feature representation with a classifier.

Baseline Methods: As baselines, we experiment with three broad representations.

Char n-grams: It is the state-of-the-art method which uses character n-grams for hate speech detection.

3)Sweeta Agrawal and Amit Awekar, 2018, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms". In Proceedings of the European Conference in Information Retrieval (ECIR). Grenoble, France: -

Past works on cyberbullying detection have at least one of the following three bottlenecks. First (Bottleneck B1), they target only one particular social media platform. How these methods perform across other SMPs is unknown. Second (Bottleneck B2), they address only one topic of cyberbullying such as racism, and sexism. Depending on the topic, vocabulary and nature of cyberbullying changes. These models are not flexible in accommodating changes in the definition of cyberbullying. Third (Bottleneck B3), they rely on carefully handcrafted features such as swear word list and POS tagging.

However, these handcrafted features are not robust against variations in writing style. In contrast to existing bottlenecks, this work targets three different types of social networks (Form spring: A Q&A forum, Twitter: microblogging, and Wikipedia: collaborative knowledge repository) for three topics of cyberbullying (personal attack, racism, and sexism) without doing any explicit feature engineering by developing deep learning-based models along with transfer learning. They have experimented with diverse traditional machine learning models (logistic regression, support vector machine, random forest, naive Bayes) and deep neural network models (CNN, LSTM, BLSTM, BLSTM with Attention) using variety of representation methods for words (bag of character n-gram, bag of word unigram, GloVe embeddings, SSWE embeddings).

3. CONCLUSIONS

The world is now more aware of the problem of spreading hate through mediums like social networks. Hate speech is a difficult phenomenon to define and is not monolithic. Every company like Facebook and twitter are working hard in regulating and countering such bad content. This attention has raised the need for automating the detection of abusive words. In this paper we analyzed the content of a post (e.g. twitter). We created dataset out of that and compared it with abusive words dataset.

Moreover, our results show the amount of abusive content that was present. In addition, this model will help in drawing road map and blueprint for future model. The future work will include sentiment analysis. The dataset we collected will be used for the perfection of training of neural network. We must also study more closely the people who use hate speech, focusing both on their individual characteristics and motivations and on the social structures they are embedded in.

REFERENCES

[1] Y. Chen, Y. Zhou, S. Zhu, and H. 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), IEEE, 2012M.

[2] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017), "Deep learning for hate speech detection in tweets", Proceedings of the 26th International Conference on World Wide Web Companion, April-3-2017.

[3] Sweeta Agrawal and Amit Awekar, 2018, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms", In Proceedings of the European Conference in Information Retrieval (ECIR), Grenoble, France.

[4] A. Joulin, E. Grave, P. Bojanowski & T. Mikolov (2016), "Bag of tricks for efficient text classification", Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2

[5] C. Nobata, J. Tetreault, A. Thomas, Y, Mehbad & Yi Chang (2016), "Abusive Language Detection in Online User Content", International World Wide Web Conference Committee (IW3C2-2016).

[6] J. Ruppenhofer & T. Kleinbauer (2019), "Detection of Abusive Language: The Problem of Biased Datasets", Proceedings to NAACL-HLT 2019