Final Project Report

Course Title: Compiler Design & Lab
Course Code: CSE332

Submitted to

Johora Akter Polin
Lecturer at Daffodil International university

Submitted by

1. Tasmima Akter
   192-15-13057
2. Tamanna Akter Swarna
   192-15-13061
3. Mayesha Iqbal
   193-15-13440

Slang And Offensive Words Detection Using Machine Learning Algorithms Based on Big Data

Tasmima Akter
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
tasmima15-13057@diu.edu.bd

Tamanna Akter Swarna
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
tamanna15-13061@diu.edu.bd

Mayesha Iqbal
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
mayesha15-13440@diu.edu.bd

***Abstract*—Negative words are quite unpleasant. On social media platforms, there are so many words that are not acceptable to everyone. Sometimes, this offensive slang badly impacts one's life which is very harmful. Initially identifying the offensive and slang words among common use comments on social media platforms like Facebook, Instagram, Twitter, and so on. The main goal of this study is to identify those words, which is not acceptable to the general public. Using the Pyspark framework and three machine learning algorithms, Logistic Regression provides the best performance of the other two. In this study, customer datasets are collected manually from different social media network sites. Thus, this study will attempt to assist in reducing the use of slang and offensive words on social media platforms.**

Keywords—Pyspark, Logistic Regression, custom dataset, slang, offensive, social media

I. INTRODUCTION

Due to the evolution of the Internet, people's communication system has changed. Nowadays, everyone's life is dependent on the internet and telecommunications technologies. People are now addicted to social media platforms. Social networking sites like Twitter, Facebook, LinkedIn, Instagram, and others have emerged as some of today's most popular places to spend time. Social media platforms enable users to engage with the global community at the touch of a fingertip and discuss their accomplishments and forthcoming ambitions. However, this technological revolution also has certain negative aspects [4]. On social media, users occasionally use various negative and offensive words when sharing their information. These negative words or comments adversely affect those who are victims of cyberbullying. These targeted victims suffer from depression, stress, and other psychological problems that can in some cases affect them so severely that they end their lives. In addition, their families often face various difficulties with these negative comments. To address the aforementioned issues, this study developed an automated model using a custom dataset manually collected from various social media platforms. In this study, we focused on Facebook, and Instagram for our data collection which has lately received a lot of attention as sources for big data research.

This paper's primary objective is to identify slang and offensive comments or words based on machine-learning approaches. Machine learning algorithms provide advantages for prediction-based research. Especially, it has been applied to the predictive model for detection from large datasets. Multiple studies of the detection of slang or offensive words that conducted over the past few years using different machine learning (ML) and deep learning (DL) based algorithms. In this particular study, we can analyze several machine-learning-based approaches applied to our unique dataset and classify slang or offensive, or good words based on the prediction models. Then, the model with the highest achieved accuracy is proposed and the influences of using the model are analyzed after validating it over the different words.

II. LITERATURE REVIEW

To predict cyberbullying and its risk factors in social media Tae-Min Song et al. conduct a prediction model using decision tree analysis in

Korea. A total of 435,563 cases of cyberbullying-related topics were used which were collected from 227 online different channels, such as news websites, blogs, online groups, social network services, and online bulletin boards. Here cyberbullying is used as a term for cyber violence. Cyberbullying-related topics were collected using 33 synonyms. Using the opinion-mining method and decision tree analysis, the types of cyberbullying were sorted using SPSS 25.0. The results indicated that the total rate of types of cyberbullying in Korea was 44%, which consisted of 32.3% victims, 6.4% perpetrators, and 5.3% bystanders [1]. In social media, most of the time words are used in different abbreviated forms, for this reason, Alok Ranjan Pal et al. used a semi-supervised learning approach to detect slang and suspicious words. In four ways slang words were handled in this study. They used supervised, and semi-supervised learning, synset, and concept analysis to detect slang, jargon, and suspicious words. Studies showed that synset and concept analysis successfully detected jargon words [2].

For sentiment analysis of Twitter data Mudassir Khan et al. used a deep learning-based RNN approach and Hadoop framework. This study tried to identify an opinion as positive, negative, or natural [3]. The proposed method offered improved classification accuracy of 0.9302, better sensitivity of 0.9404, and high specificity of 0.9157, respectively. Ghulam Murtaza et al. highlighted a new means of demonstrating aggressive behavior on SM websites. It comprehensively reviewed cyberbullying prediction models and identify the main issues related to the construction of cyberbullying prediction models in SM. This paper provides insights into the overall process for cyberbullying detection and most importantly overviews the methodology [4].

Sheetal Kusal et al. used AI-based big data sentiment analysis to detect textual big data. This study has considered 827 Scopus and 83 Web of Science research papers from the years 2005–2020 for analysis. Machine Learning based models were used to classify emotion classes. The qualitative review represents different emotion models, datasets, algorithms, and application domains of text-based emotion detection. The quantitative bibliometric review of contributions presents research details such as publications, volume, co-authorship networks, citation analysis, and demographic research distribution. The study tried to provide future research directions in this area [5]. Kazuyuki Matsumoto et al. conducted a topic analysis experiment chronologically by using the sequential Tweet data and discussing the difference in topic change according to the slang types. The study tried to analyze topics related to slang on Twitter using NLP. The Twitter data were collected which were in the Japanese language. For sentiment analysis, LSA, PLSA, and LDA methods were used in this study [6]. S. Uma Maheswari et al. tried to analyze the reviews of Social Media Big Data of E-Commerce products. To perform sentiment analysis, they used NLP which is a Machine Learning approach. Machine Learning classification models namely Multinomial Naïve Bayes, Support Vector Machine, Decision Tree Classifier, and, Random Forest Classifier was used for sentiment classification. And this sentiment classification was used as a decision support system for the customers and also for the business [7].

III. Methodology

The methodological segment of our conducted research is demonstrated hereby. This part consists of detailed information about the procedure of our dataset preparation, preprocessing datasets, and visualization of some samples of the dataset along with the process of these. The classification problem which we attempt to solve through our proposed model is also talked over in this part of the paper. The models we are using for classification and the layers and parameters of the models are determined in consideration of the dataset we have prepared for the highest accuracy and least loss.

A. Data Collection

Custom-made social media (Facebook) comments data is collected to use in this problem to predict offensive and slang words. In social media, people often use slang and offensive words. To prepare the dataset the words are collected from different comments of different people in different groups or in Ids or different accounts manually. The dataset has two attributes which are “Words” and “label” and “label” is the target attribute. The dataset contains 3880 records of data or words composed of two types of data one is slang or offensive which contains 1958 data and the other is natural words which contains 1922 data.

B. Data preprocessing

Big data and machine learning methods are used together in this problem. Since machine learning algorithm models cannot use string-type data, we had to convert the "word" attribute to a numeric value (the "word" attribute only contains slang or

offensive words that are string data). For this, we used the "Tokenizer" feature imported from PySpark and machine learning libraries. The "Tokenizer" property converts each word to a token. The dataset is then converted to feature (numeric) data using the HashingTF method because the machine learning model only accepts feature vectors. Then we need to reconstruct the dataset and the final dataset contains three features which are "word", "feature" and "label". After pre-processing we split the data in terms of train data and test data. The ratio of train and test data is 70:30 where 70% of data is used for training our model and 30% data for testing the model.

C. *Applied Model:*

For predicting the slang words different classification model is used in this paper which is the Logistic Regression Classifier, Random Forest Classifier, and Decision Tree Classifier. First of all, we train the models using final data, and the “feature” column is used as the input column and the “label” column is used as the output column. After training the models we transform all the models for prediction. After transformation, each model provided the “prediction” column which contains the predicted value of each record. For accuracy checking, we compared the “label” column and the “prediction” column. A short description of these algorithms is given as follows:

- Decision Tree

  PySpark MLlib API provides a Decision Tree Classifier model to implement classification with the decision tree method. A decision tree method is one of the well-known and powerful supervised machine learning algorithms that can be used for classification and regression tasks. It is a tree-like, top-down flow learning method to extract rules from the training data. The branches of the tree are based on certain decision outcomes. After training the model with our custom dataset, it obtained an accuracy: of 51%.

- Random Forest Tree

  A supervised learning technique built on classification tree learners is known as a random forest model. The approach generates a set of decision trees and produces a final result from all outputs. Each tree casts a vote, and the forest as a whole decides to depend on the results. The strength of each tree and the correlation of the trees determine the outcome of a vote. To perform classification using the random forest technique, the PySpark MLlib API offers the Random Forest Classifier class. After training using final data Random Forest Classifier was able to predict the test dataset 53% accurately.

- Logistic Regression

  One of the most often used Machine Learning algorithms, within the category of Supervised Learning, is logistic regression. Using a predetermined set of independent factors is used to predict the categorical dependent variable. In a categorical dependent variable, the output is predicted via logistic regression. As a result, the result must be a discrete or categorical value. Rather than providing the precise values of 0 and 1, it provides the probabilistic values that fall between 0 and 1. It can be either Yes or No, 0 or 1, true or false, etc. After training and testing it provides the highest accuracy with 98%.

IV. Evaluating the Result

We performed three different machine-learning algorithms on our custom-made dataset to correctly predict slang and offensive words. Decision Tree Classifier, Random Forest Classifier, and Logistic Regression Classifier have all been trained and evaluated on our dataset.

TABLE I. Result Evaluation

Models	**label**	**Precision**	**Recall**	**F-1 Score**	**Accuracy**
Decision Tree Classifier	0	0.50	1	0.67	51%
	1	1	0.05	0.09	
Random Forest Classifier	0	0.97	0.05	0.09	53%
	1	0.52	0.10	0.68	
Logistic Regression Classifier	0	0.97	1.0	0.98	98%
	1	1.0	0.97	0.98	

By using it across our particular dataset. It is clearly shown that Logistic Regression provides the highest accuracy which is 98% while the test loss is 0.02 which is the lowest. The test accuracy and test loss of the three ML models are shown in the Table algorithms which are Decision Tree, Random Forest, and Logistic Regression Classifiers. All the classification results are shown in Table 1 based on precision, recall, and f1-score.

A. Generating Confusion matrix

The confusion matrix visualizes the accuracy of a classifier by comparing the actual and predicted classes. The Confusion Matrix is a useful machine learning method that allows you to measure Recall, Precision, Accuracy, and AUC-ROC curve. It is denoted by four types of value that are terms True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

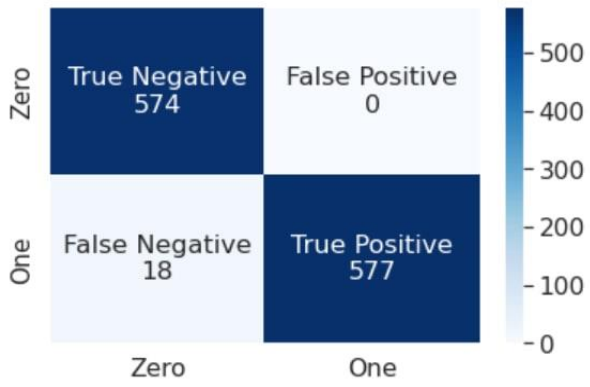

Figure 1: Heatmap of Confusion Matrix

After successfully running all the algorithms, Logistic regression Performs the best than the other two. So we may create a confusion matrix based on Logistic Regression's prediction. which will provide a summary of the algorithm's overall performance.

We may calculate the precision and accuracy of the model using these four components from the confusion matrix table. The implementation of Logistic Regression with our dataset resulted in 515 and 515 correct detections of zero and one slang and offensive words chronologically. The confusion matrix is shown in figure 1.

V. CONCLUSION

To identify slang and offensive word from social media platforms, machine learning-based algorithms are quite known. People are now addicted to social media platforms. It's very important to distinguish between negative and positive words in recent times. ML-based algorithms find slang and offensive words that are barely visible. Identify slang words among the natural words using spark framework and machine learning. Real-time data analytics using the Spark framework in a distributed computing environment. When compared to other existing methods, the proposed Spark-based machine learning algorithm of the Logistic Regression Classifier provided 98% accuracy on our custom dataset containing 3880 words. The size of the data set was not large which is considered a limitation. In a further study, the dataset will be enlarged and include paragraphs or sentences.

VI. REFERENCE

[1] Song, Tae-Min, and Juyoung Song. "Prediction of risk factors of cyberbullying-related words in Korea: Application of data mining using social big data." Telematics and Informatics 58 (2021): 101524.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Pal, Alok Ranjan, and Diganta Saha. "Detection of slang words in e-data using semi-supervised learning." arXiv preprint arXiv:1702.04241 (2015)..

[3] Khan, Mudassir, and Aadarsh Malviya. "Big data approach for sentiment analysis of twitter data using Hadoop framework and deep learning." In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1-5. IEEE, 2020.

[4] Al-Garadi, Mohammed Ali, Mohammad Rashid Hussain, Nawsher Khan, Ghulam Murtaza, Henry Friday Nweke, Ihsan Ali, Ghulam Mujtaba, Haruna Chiroma, Hasan Ali Khattak, and Abdullah Gani. "Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges." IEEE Access 7 (2019): 70701-70718.

[5] T Kusal, Sheetal, Shruti Patil, Ketan Kotecha, Rajanikanth Aluvalu, and Vijayakumar Varadarajan. "Ai based emotion detection for textual big data: Techniques and contribution." Big Data and Cognitive Computing 5, no. 3 (2021): 43.ansl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[6] Matsumoto, Kazuyuki, Minoru Yoshida, Seiji Tsuchiya, Kenji Kita, and Fuji Ren. "Slang analysis based on variant information extraction focusing on the time series topics." Int. J. Adv. Intell 8, no. 1 (2016): 84-98.

[7] Maheswari, S. Uma, and S. S. Dhenakaran. "Sentiment analysis on social media big data with multiple tweet words." International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN (2019): 2278