Cyberbullying Detection using Explainable AI with Pyspark

Salman Sadat Nur , S. M. Niaz Morshed , Ibrahim Lokman ,
Md. Mustakin Alam , Md Sabbir Hossain , and Annajiat Alim Rasel

Department of Computer Science and Engineering
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{*salman.sadat.nur, sm.niaz.morshed, ibrahim.lokman*
md.mustakin.alam, md.sabbir.hossain1}*@g.bracu.ac.bd, annajiat@gmail.com*

***Abstract*—People nowadays are more and more accustomed to virtual life on social media platforms than the real world they live in. This gave rise for the need to maintaining these platforms to keep it clean from any kind of hate speech or trying to attack someone based on their race, religion, ethnicity or something else. Monitoring and executing these manually can be really strenuous. Deployment of artificial intelligence (AI) models to do such tasks is the most effective way to monitor these platforms, especially the ones that provides explanations behind their reached conclusion. In this project, using the Pyspark environment which is a distributed computing framework, we utilized the logistic regression to create a model that is able to successfully detect a cyberbullying text with 85% accuracy. We used a data-set from a Kaggle competition consisting of 47680 data with multiple categorized cyberbullying labels to train and test the model. Later, We utilized Explainable Artificial Intelligence (XAI) to get a better understandings behind the reason of this model's reached conclusion.**

***Index Terms*—Natural Language Processing, LIME, Explainable AI, Cyberbullying Classification, Machine Learning, Text Classification, Apache Spark, Pyspark, Clustering**

I. Introduction

Cyberbullying is a significant issue people face on the internet. Attacking someone specific in the cyberspace like social media platforms based on their race, religion, age, ethnicity or any other reason is considered cyberbullying. With the growing popularity and necessity of social media hate speech and cyberbullying has also grown drastically. These leads to mental breakdowns, mental health issues and sometimes people commits suicide. However, cyberbullying can be prevented if proper steps are taken. Monitoring is a responsibility social media sites must carry on in order to keep their sites clean and healthy for their users. However, with millions and millions of concurrent active users and trillions of real-time tweets and texts publishing each minute it is not possible to monitor all of these using regular moderators and administrators. But, with modern technologies at hand AI can be very useful and efficient in monitoring and detecting cyberbullying in the cyberspace. Since internet is the archive of tremendous amount of data it is possible to train an AI to detect Cyberbullying or cyberharassment to a degree of accuracy that is satisfying.

Since, cyberbullying is a crime, it is necessary to maintain transparency and fairness in detecting such texts and Explainable AI or XAI is known for its transparency and fairness. XAI is a prominent analysis AI tool in Natural Language Processing (NLP) because unlike other AI models it is human understandable. So, when detecting a text it will show why it considers a text harassment. In this system, I used logistic regression to decide whether a text is really cyberbullying or not cyberbullying by using Explainable AI with Pyspark. Explainable AI is a set of techniques and processes that permits users to realize and rely on the outcomes and output produced by machine learning algorithms. By using it, we can debug and enhance model performance and assist others in understanding this models' behavior. Explainable AI describes an AI model with potential biases, and expected effects. It is famous for characterizing model accuracy, transparency, fairness, and results through AI-powered decision-making. XAI is needed if humans want to understand the AI-generated results and algorithms' decision reliability and organize the data in a proper way. This also helps to prevent errors in such situations where there is no scope for them. Therefore, explainable AI is a new domain of artificial intelligence where we can get answers to "why" questions that are not possible traditionally. Nowadays, XAI is used in healthcare, law, defense, and so on.

Furthermore, interpretability procedures such as Local Interpretable Model Agnostic Explanation (LIME) can assist us in choosing the ideal models. In order to do this identification and classification of cyberbullying texts, we used a data set. To predict the probability of each type of cyberbullying for individual comments, we developed a model using Explainable AI. We will analyze the classic solution using Logistic regression for classification, as Ill as Explainable AI methods.

Hence, purpose of this report is to implement a cyberbullying text detection system using Lime framework of XAI which will detect if the provided texts indicate cyberbullying or not. Rest of the report will be discussed as follows: Section II Related works on the said topic will be discussed, Section III,

Methodology of the proposed system including data-set, preprocessing, data-split will be discussed. Result analysis will be done in Section IV and finally conclusion and future work will be discussed in Section V.

II. Related Work

Detection of cyberbullying text is a well-known study, particularly in social media comments and tweets. It is challenging to detect such texts that can contain insults, vulgar words, and threats. The text-matching method- is the most traditional and widely used method to discover such text. Some of the more advanced techniques include intelligent tag-based approaches like genetic algorithms and neural networks. Using AI to detect cyberbullying in social media is also gaining popularity. Because it is much more efficient and accurate.

Ting, et. al. proposed an approach based on social media analysis and data mining for the detection of cyberbullying. Their approach contains three main techniques which include keyword matching, opinion mining, and social network analysis[1]. In another research Banerjee et. al. addressed the issue using a deep neural network where Convolution Neural Network (CNN) was utilized[2]. Hamza Wan Ali et. al. on the other hand discussed several popular approaches like Natural Language Processing (NLP) and machine learning for bullying identification [3]. Furthermore, they also discussed the issues and challenges related to cyberbullying detection. In a recent research, Khan and Bhat discussed several machine learning and NPL methods for cyberbullying detection[4]. Their study shows most of cases tf-idf was used for feature extraction. Matomela and Henney developed a system using several machine learning algorithms to detect cyberbullying on social media that makes threat in isiXhosa [5].

In this paper, we will explore the classic solution using Logistic regression for classification and more advanced techniques using Explainable AI. We aim to compare the accuracy using Explainable AI under the robust word embedding approach that could dramatically increase the accuracy score.

III. Spark Environment

Apache Spark is a dynamic tool of Big Data and an open-source distributed cluster-computing framework. It incorporates a common Machine Learning (ML) library (MLlib) designed to ML classifiers. Spark functions well in memory and supports many programming languages such as java, python, SQL, scala, R) and works well with Python. Spark ecosystem is composed of Spark core component, Spark SQL, Spark Streaming, Spark MLlib, Spark GraphX, and SparkR. In this project, we use Spark MLlib. MLlib is a scalable Machine Learning library that contains Machine Learning libraries with the implementation of various Machine Learning Algorithms. One of the key features of Apache Spark is its rapid processing. Apache Spark exhibits high data processing swiftness(about 100x faster in memory and 10x faster on the disk). Another important feature is that Apache Spark is highly dynamic. A parallel application can be developed in Spark by the user. By using Spark we can better the processing speed in-memory. Re-usability is another key characteristic, users can easily reuse spark code for batch processing.

IV. METHODOLOGY

The steps in the methodology are as follows shown in the Fig 1

A. Dataset

The dataset we used in this project was provided by J. Wang et. al. in a Kaggle competition [6]. It includes tweet texts from Twitter that falls under five categories of bullying. They are gender, age, religion, ethnicity, and other_threat. It was a CSV (comma-separated values) formatted dataset. Each row of a CSV file corresponds to one row of the dataset. The PySpark MLlib library was utilized to work with this dataset after converting it into data frame. More specifically, we utilized the train.csv file to train and validate this model, which has 47680 rows and 8 columns, the first row of which is the header. One remark may relate to several categories since the dataset has several labels. For example, according to our study, a text may be religious, age, and ethnicity related cyberbullying all at once.

B. Preprocessing

Online tweets are mostly non-standard English consisting of emojis, typos, non-conventional trendy misspellings of a word & case mismatching. It is difficult to work with such text data and the model accuracy could drastically fall if the model is trained with such data. For this reason, We defined a function named “clean_text” to clean the data set’s misspelling ’tweet_text’ column texts before using them to train this model. After that, we introduced a new column in the data set named ’Cyberbullying/Not-cyberbullying’ which indicates 1 if a text falls under any one of the five labeled categorized cyberbullying, as a text can fall under multiple categories at a time and 0 is indicated when it falls under none of the five labeled categorized cyberbullying. We also defined another function named “remove_stop_words” to remove some specific stop-words which we mentioned in a list. These are pronouns and words which are not supposed to be categorized.

C. Data Split

We have divided the pre-processed training data into 2 sets to train & validate our model utilizing sci-kit learn.

- Training Data: The dataset contains 80% of the total data upon which the model was trained on.
- Testing Data: The dataset contains 20% of the total data upon which the model was tested on.

D. Creating Word embedding and vectorization

It is crucial for any NLP task to do effective word embeddings by transforming texts into matrix representations of numbers in order for any algorithm to make meaning out of the texts from any dataset. We used an unsupervised weighting scheme named TF-IDF (Term Frequency - Inverse Document

Fig. 1. Workflow Diagram

Frequency) to create word embeddings consisting sparse matrix. This is done to enable a model to identify how relevant a word is in a document by fitting and transforming the text data into vector representations. In code, we imported the TfidfVectorizer from sklearn library's 'feature_extraction.text' package to do this task.

E. Defining Classifier

We used logistic regression classifier and later evaluated its performance. Logistic regression is a statistical model that predicts binary output based on prior inspection. This algorithm predicts a dependent outcome by computing the relationship between the other existing relevant independent variables. Since our task is to classify a text being cyberbullying or not, We utilized the Logistic Regression from sklean library's "linear_model" package by setting the hyperparameter values C = 5.0, penalty='l2', solver = "liblinear" & random_state=45.

V. Result Analysis

The following are the stages involved in the result analysis:

A. Model Evaluation and Comparison

By dividing the dataset 75% training data, 25% test data, We have run our experiment. Logistic regression showed accuracy of 85%. The classifier showed, 79.33% were true positive, 11.33% false positive, 5.34% true negative and 4.01% false negative results.

We compared our model with two other machine learning algorithm model Linear Support Vector Machine(SVC) and Bermoulli's Naive Bayes. The Table I shows the performance of the classifiers. Where Linear SVC and Logistic Regression have the highest accuracy of 85% and Bermouli's Naive Bayes predicts 79% accurately.

TABLE I
COMPARISONS OF OTHER ALGORITHMS WITH OUR MODEL

Method	Class	Precision	Recall	F1-score	Accuracy
Linear SVC	Not-Cyberbullying	0.59	0.36	0.44	0.85
	Cyberbullying	0.88	0.95	0.91	
Logistic Regression	Not-Cyberbullying	0.57	0.31	0.40	0.85
	Cyberbullying	0.87	0.95	0.91	
Bernoulli Naive Bayes	Not-Cyberbullying	0.42	0.77	0.55	0.79
	Cyberbullying	0.94	0.79	0.86	

The table I shows the performance of our classifier on the test set consisting of 9536 samples. Also, generated confusion matrices for the classifier that is shown below in Fig 2 for Logistic Regression. The confusion matrices in Fig 2 is showing that the classifiers are successfully predicting the correct classes for almost all of the samples from the testing set consisting of 11920 samples. Logistic Regression classified 85% of the samples correctly from the training set.

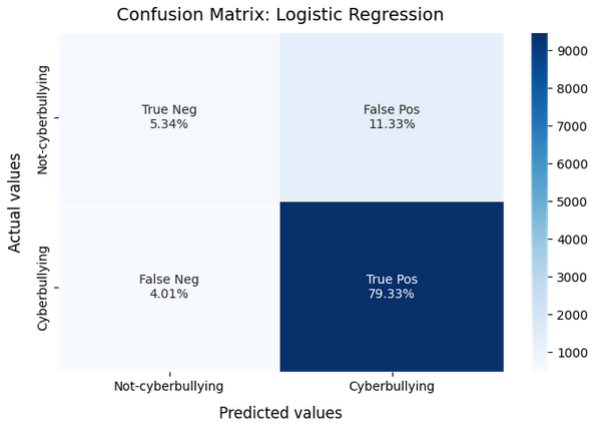

Fig. 2. Confusion Matrix for Logistic Regression

We also generated the precision, recall & f1 score for both of the two classes (Not-cyberbullying & Cyberbullying) using these data for all three classifiers that are given in the table I.

B. Result Analysis with XAI

In this context XAI is used to get a better understanding about the model's output and to understand the reason behind a model's reached conclusion. We utilized LIME as our XAI framework. This gives us insights behind our output and the logic behind providing that particular output based on the words in the sentence. Few samples are provided below. We implemented LIME only on the logistic regression's output.

Let it be mentioned that there are three sections in the output, the first section of the output is the resultant label and the fraction, the second part is list of words and their contribution percentage to the classifier's generated output finally the third section is the entire text and the highlighted words based upon which the classifier is assigning a label to the final output.

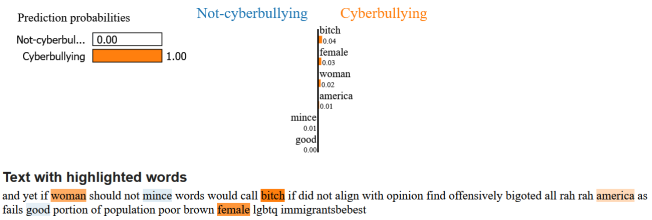

Fig. 3. Lime method for the sample number 3829

In the Fig. 3, it has classified the sample number 3829 as a cyberbullying text by analyzing the words of that sentence. A lot of slang and hate comment related words has been used here in the text and the classifier classified it as a cyberbullying comment based on those words We can see that LIME is classifying this text as a 100% cyberbullying text and highlighted all the obscene words based upon which the classifier is labelling the output.

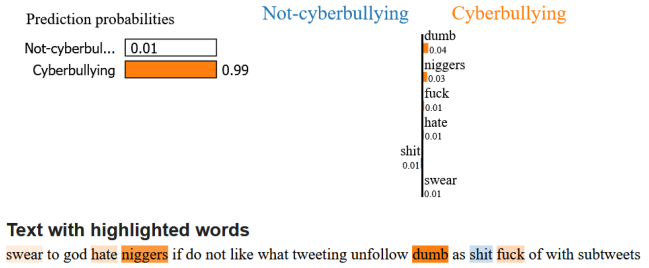

Fig. 4. Lime method for the sample number 1535

Similarly, in Fig. 4 We can see that there are multiple cyberbullying words based upon which it is being labelled cyberbullying. Since most of the words are labelled as cyberbullying LIME is classifying this text 99% cyberbullying text.

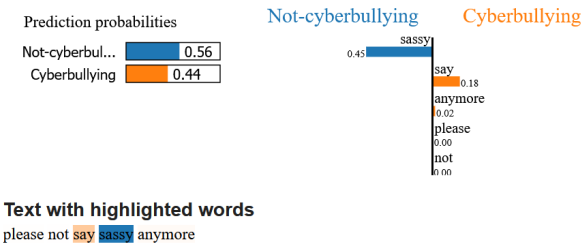

Fig. 5. Lime method for the sample number 215

In the Fig. 5, it has classified the sample number 215 as a Not-cyberbullying text based on the consisting words of that sentence. The word ‘sassy’ is carrying most of the weight behind the prediction of the whole sentence’s sentiment being positive. It is due to it being labeled as a positive word in the data-set that was used to train the model. The word ’say’ is not exactly a bullying word but it is being labelled as cyberbullying due to being trained on a data-set where this word had a cyberbullying tag. We can see that LIME is classifying this text as a 56% not-cyberbullying text and highlighted all the words based upon which the classifier is labelling the output.

VI. Conclusion and Future Work

In order to provide a safe virtual atmosphere, it is very important to filter out cyberbullying comments on any social media platform. Several traditional machine learning approaches are still available that are being used to detect cyberbullying comments, but their performance and accuracy is not good enough for the growing number of users and the large data-set of comments. We used explainable AI to detect and classify social media comments into cyberbullying and not-cyberbullying texts. As shown in the previous sections, Explainable AI provided interpretability and explanation of the decisions made. This system focused on detecting cyberbullying comments with high accuracy. It showed why it decided the tweets were cyberbullying-related or not. Explainable AI used in this system has shown an accuracy of 84%. However, more testing needs to be done on real-time data and further experimentation might reveal any weakness of the system.

There is still room for improvement in this system. One of the main concerns is to reduce the false negative rate. In order to do that we will integrate syntactic features on top of the current feature set which will help to state the tweet text if someone is cyberbullying or not. Although the system showed good accuracy and Explainable AI provides transparency and fairness in detecting cyberbullying by declaring the reasons behind the decision it took. We also want to integrate sentiment analysis in the future. So that it is possible to better understand the sentiment behind these tweets. So that it is possible for the platform and the authorities to take proper steps to stop these kind of behaviour and activities in these platforms and reduce the growing number of mental health issues and social media bullying. Not only texts, people also use emojis in social media comments to bully people. So, integration of a detection system that detects misuse of emoji is another of my goals.

References

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

[2] Vijay Banerjee, Jui Telavane, Pooja Gaikwad, and Pallavi Vartak. Detection of cyberbullying using deep neural network. In *2019 5th International Conference on Advanced Computing Communication Systems (ICACCS)*, pages 604–607, 2019.

[3] W. N. Hamiza Wan Ali, M. Mohd, and F Fauzi. Cyberbullying detection: An overview. In *2018 Cyber Resilience Conference (CRC)*, pages 1–3, 2018.

[4] Asif Ahmad Khan and Aruna Bhat. A study on automatic detection of cyberbullying using machine learning. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1167–1174, 2022.

[5] Vuyokazi Matomela and Andre j. Henney. Cyberbullying detection system focusing on the isixhosa language. In *2022 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–6, 2022.

[6] k Wang, J.and Fu and C.T Lu. Cyberbullying classification. In *Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020)*, pages 10–13, 2013.