Toxic Comment Detection using Explainable AI

1st Shafin Mahmud Jalal
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
shafin.mahmud.jalal@g.bracu.ac.bd

2nd Md. Rezuwan Hassan
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.rezuwan.hassan@g.bracu.ac.bd

3rd Rufaida Tasin
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
rufaida.tasin.@g.bracu.ac.bd

4th Masum Uddin Ahmed
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
masum.uddin.ahmed@g.bracu.ac.bd

5th S.M Niaz Morshed
Computer Science and Enginnering
BRAC University
Dhaka, Bangladesh
sm.niaz.morshed@g.bracu.ac.b

6th Md Humaion Kabir Mehedi
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

7th Md Mustakin Alam
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.mustakin.alam@g.bracu.ac.bd

8th Annajiat Alim Rasel
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—People nowadays are relying more and more on social platforms for communication for it's easy and wide accessibility. This gave rise to maintaining these platforms to keep it clean from any kind of hate speech or any sort of statements consisting any vulgar or offensive words. Monitoring executing these manually can be really strenuous. Deployment of 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 paper, We utilized the Logistic Regression to make a model that is able to successfully detect a toxic comment with 96% accuracy. We used a dataset from a Kaggle competition consisting of 159570 data with multiple categorized toxicity labels to train and test our model. Later, We utilized Explainable AI (XAI) to get a better understandings behind the reason of our model's reached conclusion.**

***Index Terms*—LIME, Explainable AI, Toxic comment classification, Machine Learning, Text Classification**

I. Introduction

Toxicity in the internet recently is a significant issue people face when they post some text publicly on the internet about their life and share pictures, videos, thoughts, and so on. In the case of teasers, trolls, and scammers, those who share posts and comment on unsuitable and abusive subjects on social sites are increasing sharply. Multiple cases are found due to this online toxicity; numerous users are bound to make their social media profiles inactive or mute the comments section as if no one could put comments on their images or posts. Besides these, this consequence has been responsible for making severe cases that have even been responsible for losing innocent life. Victims of cyberbullying and online harassment cultivate to grow less confidence, lower self-esteem, and increased suicidal perceptions. Content Moderation is a method to control user-submitted content based on some specific rules and policies to decide whether a submission is tolerated or not. As a result, Content moderation is vital to maintaining a safe and comfortable online experience since it helps in saving the users from being the victims of problematic and harsh content. It is challenging to detect such toxic texts that can contain insults, vulgar words, and threats. The text matching method is the most traditional and widely used method to discover such toxic text. In a single day, many comments and tweets are published online, so tracking and flagging problematic and harmful words by the help of moderators becomes inadequate. In terms of this scenario, automated moderation can be the key to solving this issue by flagging or blocking the occurrences of toxic and inappropriate comments automatically. We know that machine learning models can now predict new and unseen data. Nevertheless, these models do not explain how distinct features contribute to the output. So, AI is called a black box for its functioning behavior. For this, It may not deliver explanations in complex scenarios such as diagnosing in defense or life-threatening diseases. If a description/Explanation, along with the output result, is combined with human reasoning, it may establish output beneficially. This scenario creates the foundation of explainable artificial intelligence (XAI). XAI provides explanations to many queries along with the Result. Explainable AI is becoming a prominent analysis area and has discovered many applications in different areas. In this study, we checked the performance of different deep learning and machine learning models to decide whether a word is really toxic or non-toxic by using Explainable AI. 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 our models' behavior. Explainable AI describes an AI model, potential biases, and expected effect. It is famous to characterize model accuracy, transparency, fairness and results through AI-powered decision-making. XAI is needed if humans want to understand the AI results, and algorithms' decision reliability, and organize the data in a proper way. Explainable AI builds trust in the AI regulators and business associates for commercially valuable and for the right decision-making. In the medical field, when XAI will be used the result with explanations helps to find the exact reasoning behind diseases. 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 which is not possible traditionally. Nowadays, XAI is used in healthcare, law, defense, and so on. Also, we find that the best preprocessing method and models will be the perfect combination for each other. Furthermore, interpretability procedures such as LIME can assist us in choosing the ideal models. In order to do this identification and classification of toxic comments, we used a dataset. To predict the probability of each type of toxicity for individual comments, we developed a model using Explainable AI. We will analyze the classic solution using Logistic regression for classification, as well as Explainable AI methods.

II. Literature Review

Classification of toxic comments is a well-known study, particularly those comments in English languages. It is challenging to detect such toxic texts that can contain insults, vulgar words, and threats. The text matching method- is the most traditional and widely used method to discover such toxic text. The application uses a database to check whether the comment matches or not with the database. This technique could not be effective because people tend to bypass this detector and create unknown words that express the same toxicity but unique spelling. Shortening the word or message also damages the database efficiency. Besides, Due to the lack of data, this classification research is not well analyzed in some countries like Indonesia. Minimizing the spread of toxicity is an actual problem on social media sites, which has received a lot of attention in recent years.Different approaches have been proposed to manage these issues. Categorizing toxic words using Convolutional Neural Network (CNN), NBSVM, and Long Short-Term Memory (LSTM).TF-IDF extraction method used by the model of NB-SVM are few of the ways to detect toxicity in order to get rid of them is one of the ways to deal with this issue. Many researches have been done on multilabel classification using several Binary Relevance classification methods such as Naïve Bayes, SVM, and KNN and several feature selections such as Chi-Square, Mutual Information, and Odd ratio. Each field uses different classification definitions; similar methods can often be applied to various tasks. A Recent research on toxic comments detection using natural language processing (NLP) techniques along with machine learning strategies [1] and [2]. In terms of identifying the text and network features, this research was successful [3]. Moreover, [4] Using deep learning techniques, toxic keyword identification, and other bullying detection were studied [5] [6] . Nowadays, Using the Explainability of artificial intelligence procedures along with deep learning and machine learning methods to comprehend the reasoning for labeling toxic comments in different social media and medical applications has increased significantly. The explainable AI method LIME (Local Interpretable Model-agnostic Explanations) was discussed well [3] [7] to make its predictions individually understandable [8], and the finest approaches for the use of these interpretable machine learning models and their applications were also discussed [9] [10] [11]. Explainable AI (XAI) is now popular in decision-making using AI techniques. Notable descriptions about explainable machine learning and deep learning [12] in the categorization of XAI techniques and strategies on different factors such as their scope, methodology, algorithmic instinct and the ability of explanation [13] XAI models are available. For example, LIME, layer-wise relevance DeepLIFT, and propagation along with their deployment [14] [15] [16]. It is used now as predictive maintenance (PdM) in production [17] and also for social science research [18]. 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. METHODOLOGY

The steps in the methodology are as follows:

A. Dataset

The dataset [?] that we have used in this project was provided by Conversation AI (founded by Jigsaw and Google) in a Kaggle competition. It includes Wikipedia comments that falls under six categories of toxicity. Like, toxic, severe toxic, obscene, threat, insult, indentity_threat. It was a CSV (comma-separated values) formatted dataset. Each row of a CSV file corresponds to one row of the dataset. The Python Pandas library was utilized to work with this dataset after converting it into dataframe. More specifically, We utilized the train.csv file to train and validate our model, which has 159570 rows and 8 columns, the first row of which is the header. One remark may relate to several different categories since the dataset has several labels. For example, according to our study, a comment may be toxic, obscene, and insult all at once.

B. Preprocessing

Online comments 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 dataset's misshapen 'comment_text' column texts before using them to train our model. After that we have introduced a new column

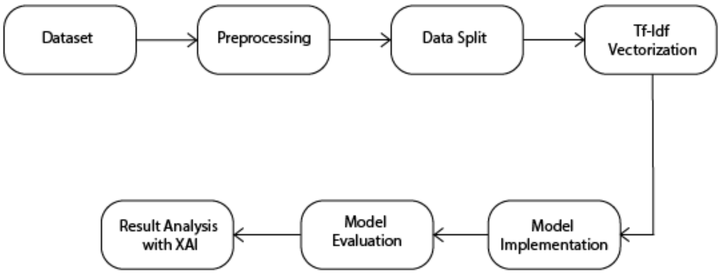

Fig. 1. Workflow Diagram

in the dataset named 'Toxic/Non-Toxic' which indicates 1 if a comment falls under any one of the six labeled categorized toxicity as a comment can fall under multiple categories at a time and 0 is indicated when it falls under none of the six labeled categorized toxicity.

C. Data Split

Machine Learning models are trained and tested on different sets of data. We have divided the pre-processed training data into 2 sets to train validate our model utilizing scikit learn.

- Training Data: The dataset upon which the model would be trained on, contains 80% data.
- Testing Data: The dataset upon which the model would be tested on, contains 20% data.

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 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.

E. Defining Classifier

Logistical regression is a statistical model that predicts binary output based on prior inspection. This algorithm predicts a dependent outcome computing the relationship between the other existing relevant independent variables. Since our task is to classify a comment being toxic 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.

IV. Result Analysis

The following are the stages involved in the result analysis:

A. Model Evaluation

With 80% training data, 20% test data, we have run our experiment.

As seen in Fig. 2, The classifier is successfully predicting the correct class for most of the comments from the testing set. It is successfully predicting 30610 comments correctly out of the 31915 comments. Which is approximately almost 96% of the comments correctly classified.

B. Classification Report

TABLE I
Classification report of our model

		Precision	Recall	F1-score	Accuracy
Class	Non toxic	0.97	0.99	0.98	96%
	Toxic	0.86	0.71	0.78	
Micro avg		0.91	0.85	0.88	
Weighted avg		0.96	0.96	0.96	

The table I shows our model's performance on the test set. It was able to correctly classify approximately almost 96% of all the comments out of all the 31915 entries. The precision, recall, f-1 score and support of the two classes (Non-Toxic Toxic) are 0.97, 0.99 , 0.9, 28670 and 0.86, 0.71 , 0.78, 3245 respectively

Also, The precision, recall and f-1 score of the two averages (Macro Weighted) are 0.91, 0.85, 0.88 and 0.96, 0.96, 0.96 respectively.

V. Future Work

There is room for improvement in our system. Our goal is to reduce the false negatives rate. In order to do that we will integrate syntactic features on top of the current feature set which will help to state the role of a toxic word in a comment if that is toxic or non-toxic. Although the system showed great accuracy and Explainable AI provides transparency and fairness in detecting toxic comments 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

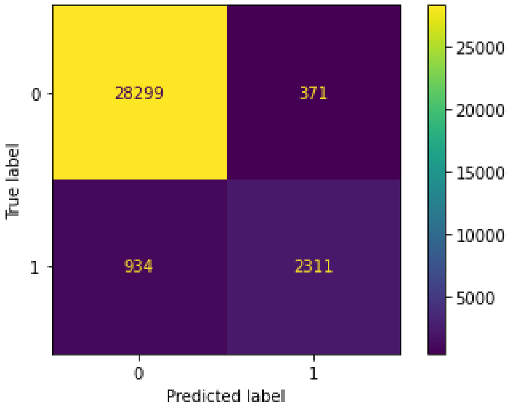

Fig. 2. Confusion Matrix

the sentiment behind these comments and the accuracy of the toxic comment detection will be further improved. Furthermore, people will try to evade this system by misspelling toxic words, or some people may misspell a word into a toxic word. So, we aim to integrate spelling and grammar correction tools in order to avoid misclassification or intentionally misspelled words. Our goal is to make this system as robust as possible and integrate it in as many social media platforms as possible. So that we may reduce the toxicity 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 our goals.

VI. Conclusion

In order to provide a safe virtual atmosphere, it is very important to filter out toxic comments in any social media platform. Several traditional machine learning approaches are still available that are being used to detect toxic comments, but their performance and accuracy is not good enough for the growing number of users and the large dataset of comments. We used explainable AI to detect and classify social media comments into toxic and non-toxic comments. As shown in the previous sections, Explainable AI provided interpretability and explanation of the decisions made and it showed better results of the given dataset. Our system focuses on detecting toxic comments with high accuracy. It showed why it decided the comments were toxic or not. Explainable AI used in this system has shown accuracy of 96%. Which outperforms other methods such as ernoulli's naive bayes method, logistic regression and Linear SVC. However, more testing needs to be done on real time data and further experimentation might reveal any weakness of the system.

References

[1] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.

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

[3] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22, 2017.

[4] Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*, pages 105–114, 2019.

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[6] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. *arXiv preprint arXiv:1706.07206*, 2017.

[7] Aditya Mahajan, Divyank Shah, and Gibraan Jafar. Explainable ai approach towards toxic comment classification. In *Emerging Technologies in Data Mining and Information Security*, pages 849–858. Springer, 2021.

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[9] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

[10] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[11] Michael Hind, Dennis Wei, Murray Campbell, Noel CF Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Ted: Teaching ai to explain its decisions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129, 2019.

[12] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[13] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

[14] Olli Kanerva et al. Evaluating explainable ai models for convolutional neural networks with proxy tasks. 2019.

[15] Harshkumar Mehta and Kalpdrum Passi. Social media hate speech detection using explainable artificial intelligence (xai). *Algorithms*, 15(8):291, 2022.

[16] Alberto Fernandez, Francisco Herrera, Oscar Cordon, Maria Jose del Jesus, and Francesco Marcelloni. Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? *IEEE Computational intelligence magazine*, 14(1):69–81, 2019.

[17] Bahrudin Hrnjica and Selver Softic. Explainable ai in manufacturing: a predictive maintenance case study. In *IFIP International Conference on Advances in Production Management Systems*, pages 66–73. Springer, 2020.

[18] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.