Pattern Recognition and Explainable AI: Racist text detection using NLP

*Note: Sub-titles are not captured in Xplore and should not be used

1st Salman Sadat Nur
CSE
BRAC UNIVERSITY
Dhaka , Bangladesh
salman.sadat.nur@g.bracu.ac.bd

2nd Labiba Ifrit Jahin
CSE
BRAC UNIVERSITY
Dhaka , Bangladesh
labiba.ifrit.jahin@g.bracu.ac.bd

3rd Tamanna Sultana Tonu
CSE
BRAC UNIVERSITY
Dhaka , Bangladesh
tamanna.sultana.tonu@g.bracu.ac.bd

4th Sushana Islam Mim
CSE
BRAC UNIVERSITY
Dhaka , Bangladesh
sushana.islam.mim@g.bracu.ac.bd

5th Golam Kibria Anim
CSE
BRAC UNIVERSITY
Dhaka , Bangladesh
golam.kibria.anim@g.bracu.ac.bd

***Abstract*—**
***Index Terms*—**

I. Introduction

Racism is a significant issue on the internet, affecting people's attitudes, responses, feelings, and opinions regarding specific events, individuals, and entities. The increasing prevalence and need for virtual entertainment has led to increased discourse and prejudice, causing mental breakdowns, mental health issues, and even suicide. To address this issue, social media platforms must monitor and maintain a clean and safe environment for their users. Interpretability methods like Local Interpretable Model Agnostic Explanation (LIME) can help select suitable models. This report aims to use the LIME framework of XAI to build a text detection system for racism that can determine if a text indicates racism. The proposed system's methodology includes emotional analysis, as Twitter feeds are often dominated by floods of emotion and affect, which can divide narratives into polarities of good and evil. Explainable AI, or XAI, is known for being honest and transparent, making it a popular analysis AI tool in Natural Language Processing (NLP). The system uses logistic regression to determine if a text is actually racism. Explainable AI is a collection of methods and procedures that enable users to understand and rely on the results and output produced by machine learning algorithms. Logical artificial intelligence (XAI) is a computer-based intelligence model that uses AI-powered decision-making to characterize model accuracy, transparency, fairness, and outcomes. It is essential for humans to understand AI-generated results, algorithmic decision reliability, and data organization. Explainable AI is a new subfield of artificial intelligence that can provide conventionally impossible answers to "why" questions. Real-time tweets and texts published daily can be used to obtain conventionally impossible answers. AI can be helpful and effective in monitoring and detecting racism in cyberspace with the help of current technologies. It is possible to train artificial intelligence to detect racism or cyber-harassment with satisfying accuracy, as the internet is a vast source of data. The report aims to use the LIME framework of XAI to develop a text detection system for racism, identifying and categorizing racist texts. The system uses a data set and Explainable AI to predict the likelihood of each type of racism for individual comments. The report covers the methodology, data set, pre-processing, and data-splitting, related works, results analysis, and conclusions, and future work. The report also discusses related works and the results of the study.

II. Literature Review

A. *Advancing XAI in Hate Speech Detection: Interpreting Complex AI Models*

This research conducted by Mehta et. al[1] aims to interpret and explain decisions made by complex artificial intelligence (AI) models in hate speech detection. Two datasets were used to demonstrate hate speech detection using XAI. Data preprocessing was performed to clean data of inconsistencies, tokenize and lemmatize the text, and simplify categorical variables for training purposes. Exploratory data analysis was performed on the datasets to uncover various patterns and insights. Various pre-existing models were applied to the Google Jigsaw dataset, such as decision trees, k-nearest neighbors, multinomial naïve Bayes, random forest, logistic regression, and long short-term memory (LSTM). LSTM achieved an accuracy of 97.6%. Explainable methods such as

LIME were applied to the HateXplain dataset, and variants of the BERT (bidirectional encoder representations from transformers) model, such as BERT + ANN and BERT + MLP, achieved good performance in terms of explainability using the ERASER benchmark. The research presents practical XAI algorithms such as occlusion, integrated gradients, and LRP, and proposes a framework called "teaching explanations for decisions (TED)" to provide explanations of an AI system. However, the research has limitations, such as the lack of a method for performing the pick-up step for images, the difficulty of interpreting explanations, and the lack of detailed explanations in semantic embeddings and lexicon expansion techniques.

B. Detecting Cyberbullying: Insights from Data Mining and Machine Learning Approaches

Ting and colleagues have presented a technique for identifying cyberbullying that is based on the study of social media and data mining. Keyword matching, opinion mining, and social network analysis [3] are the three primary methodologies used in their strategy. In a separate piece of research, Banerjee and colleagues [4] addressed the problem by using a deep neural network that was constructed using a CNN (Convolution Neural Network). And Hamiza Wan Ali al., on the other hand, addressed a variety of well-known technologies for spotting bullying, such as machine learning and Natural Language Processing (NLP) [5]. In addition to this, they discussed the challenges and difficulties associated in locating instances of cyberbullying. In a recent work [6], Khan and Bhat discussed a variety of machine learning and natural language processing strategies that may be used to identify instances of cyberbullying.According to the findings of their investigation, tf-idf was used the vast majority of the time while feature extraction was being performed. Matomela and Henney developed a machine learning-based approach that can recognise isiXhosa-language cyberbullying on social media [7].

C. GCR-NN: Superior Sentiment Analysis Model for Detecting Racist Tweets

This study made by Lee et. al [2]aims to detect racist tweets using sentiment analysis using a stacked ensemble deep learning model, Gated Convolutional Recurrent Neural Networks (GCR-NN). The model combines GRU, CNN, and RNN, extracting prominent features and CNN for accurate predictions. The results show that the GCR-NN model can detect 97% of tweets containing racist comments, demonstrating its superior performance in sentiment analysis tasks. The study emphasizes the importance of monitoring and addressing racism remarks on social media platforms, as they can lead to mental and physical health issues and contribute to the spread of racist opinions. Machine learning and deep learning approaches have proven their strength and superiority in various domains, including sentiment analysis.The study presents a large dataset of tweets containing racist comments/text and compares the performance of various machine learning models, including GRU, LSTM, CNN, and RNN. The proposed system detects hate speech across multiple social media platforms using sentiment analysis to identify racist tweets. The study aims to optimize the performance of these models by fine-tuning several hyperparameters. The proposed GCR-NN model outperforms all machine learning and deep learning models with a 0.98 accuracy score. The study suggests that automatic racism detection and stopping racist comments on social media platforms like Twitter is crucial to prevent further spread.

D. Detection of Cyberbullying in Social Media Texts Using Explainable Artificial Intelligence

The research paper worked on cyberbullying detection using machine learning techniques. It also showed content about sentiment scores, TF-IDF scores, weight scores, and bootstrap samples. The thesis also analyses several machine learning techniques for detecting cyberbullying, including random forest classifier, gradient boosting, support vector machine, and long short-term memory. In this paper, they wanted to create a brand-new method of detecting cyberbullying based on attributes like gender, religion, age, and ethnicity. Secondly, they planned to enable the proposed approach to provide justification for why a text is classified as a particular type of cyberbullying. Lastly, they wanted to increase the suggested method's categorization accuracy. Their proposed model is built into three layers: Text Pre-processing, Text Training, and Decision Explanation. In Text Pre-processing Layer they used three Natural Language Processing (NLP) techniques: text cleaning, sentiment analysis, and Term Frequency-Inverse Document Frequency (TF-IDF). Lastly, The Decision Explanation Layer utilizes Explainable Artificial Intelligence (XAI) to provide explanations for the decisions made by the machine learning models in the second layer. It implements decision trees and a random forest model to make predictions and justify the decisions. Therefore, the paper showed the use of LIME demonstrates that the proposed models provide similar justifications for their decision-making processes.

E. An Explainable Artificial Intelligence Model for Detecting Xenophobic Tweets

The research paper is based on detecting xenophobic content in tweets using machine learning techniques. The paper proposes a new feature representation for xenophobia detection based on sentiment analysis, syntactic analysis, and semantic analysis. The authors use three NLP APIs and the Python spaCy library for extracting features from tweets. Additionally, they employ a variety of machine learning algorithms and contrast pattern mining techniques. The study covers the benefits and drawbacks of black-box and white-box models while emphasizing the significance of interpretable models. The main contribution of this research is to provide an XAI model in a language close to experts in the application area, such as psychologists, sociologists, and linguists.[5] As a part of this research, they have created a Twitter database in collaboration with experts.[5] In this paper, they extracted new features using Natural Language Processing (NLP) approach. Then

they used XAI approach to create a robust and understanding model for experts. Furthermore, to classify xenophobic tweets they used a contrast pattern-based classifier. They tried to show their proposed model was more interpret-able than other approaches. Overall, the paper showed that the application of XAI and contrast pattern-based classifiers could help experts in the field of xenophobia categorization to better understand the logic for classification decisions.

III. Methodology

A. *Dataset*

We have accumulated a diverse collection of messages sourced from Twitter, encompassing various types, including a mixture that may include offensive content. To commence the analysis of the dataset, we employed Python's Pandas package. The dataset consisted of multiple labels, such as 'index,' 'id,' 'Text,' and 'Annotation.' Among these, we focused primarily on the 'Annotation' column, as it indicated whether the content was categorized as racist or non-racist. To ensure data integrity, we checked for duplicates by applying the 'duplicated' function to the 'id' column after reading the dataset from the .CSV file. Initially, our dataset comprised 13,471 rows and 5 columns, signifying that there are 13,471 rows of text.

B. *Cleaning and Preprocessing*

In order to be certain that our detection algorithms provide accurate results, it is necessary for us to clean and preprocess the text input. There are a variety of errors included in the data, including truncated phrases, emojis, misspellings, improper case matching, and contractions of several words. In order to find a solution to these issues, we developed a function that can carry out a number of different responsibilities. The first thing that we did was transform contractions into their full-blown forms. For instance, the word "can't" was changed to "cannot," and the word "won't" was changed to "will not." This process helps to standardise the language and eliminates any ambiguity that may have been generated by contractions.

The next step was to repair any misspellings by substituting the proper version of the term. In order to do this, several methods, such as spell-checking algorithms and reference dictionaries, had to be used so that misspelt words could be located and replaced with their proper counterparts. We increase the reliability of the text data as a result of doing so.

In addition to that, we stripped the words of any special characters or emojis that could have been there. These non-alphabetic symbols may not add much to the identification algorithms, and they may also have the ability to inject noise or undesirable fluctuations into the data.

However, owing to the fact that some letters and words were eliminated during this cleaning procedure, it's probable that some of the cells in the spreadsheet are now empty. In order to address this issue, we made use of the mask approach and substituted these empty cells with the text "NaN," which stands for "Not a Number" and denotes the absence of data.

After that, we used the "dropna" technique to get rid of the rows that had the "NaN" value. The number of rows in our dataset was cut down to 13,426 thanks to this operation, which means that we are now dealing with a dataset that is both clean and consistent.

In addition, we compiled a list of common "stop words," which are words with just one letter or no more than one, such as pronouns and words with only one letter. These are words that, in the context of the study, often do not have any meaningful significance. After that, we deleted these stop words from the texts, which further refined the data for the analysis that came after that.

C. *Data Split*

We carried out a data split in order to get our data ready for the training and the assessment that would follow. We choose to use the "Text" column as our feature (X), which houses the textual data, and the binary values contained in the "Annotation" column as our goal (y).

In order to do the data split, the dataset was first partitioned into two subsets: a training set and a test set. The whole of the dataset was divided into two parts: the training set, which contained 75% of the total, and the test set, which included the remaining 25%. Because the data has been partitioned in this way, we are able to train our models using a sizable chunk of the data while retaining a distinct component for analysing how well the models perform.

D. *Word Embedding and Vectorization*

Word embedding is a method that is used to convert text data into numerical representations, which enables machine learning algorithms to comprehend and process them efficiently. Vectorization is another approach that is used to translate text data into numerical representations. To generate word embeddings in the form of sparse matrices, we used the "TF-IDF" (Term Frequency-Inverse Document Frequency) approach in this specific instance.

In order to do this, we made use of the "TfidfVectorizer" class that was included inside the Sci-kit learn package. This class is responsible for implementing the TF-IDF algorithm, which determines the significance of each word in a text by taking into account both the frequency with which that word appears in the particular document and the rarity with which it appears over the whole dataset. The end result is a numerical representation of the text data that accurately conveys the importance of each individual word in relation to the whole of the dataset.

E. *Defining Classifiers*

*1) **Logistic Regression**:* When we first began developing our classification system, we relied heavily on logistic regression. Logistic Regression is a popular approach for machine learning that is often used for binary classification projects. This method determines the likelihood of a sample belonging to a certain class by first fitting a logistic function to the training data and then making a prediction based on that probability.

In order to make use of logistic regression, we imported the 'LogisticRegression' function that is included inside the

'Linear-Model' package that is located within the Sci-Kit Learn library. With the help of the tools that are provided by this function, it is possible to train a Logistic Regression model using the preprocessed text input.

*2) **Bernoulli's Naive Bayes**:* Bernoulli's Naive Bayes In addition to Logistic Regression, we compared our findings using the Naive Bayes method. This was done in place of Logistic Regression. A probabilistic classifier known as Naive Bayes employs Bayes' theorem under the assumption that the characteristics are independent of one another in order to provide a prediction about the likelihood of a sample belonging to a certain category.

To be more specific, we used Bernoulli's Naive Bayes approach, which is designed to work with binary feature vectors. In our particular scenario, the word embeddings that were generated via the use of the TF-IDF approach supplied the algorithm with the essential binary feature vectors.

We went beyond Logistic Regression in our research by using the Naive Bayes method, which allowed us to investigate the efficacy of a new strategy in relation to the current endeavour we were working on. Because of this comparison, we are able to assess and pick the method that will be the most successful in solving our text categorization issue.

IV. Result Analysis

Performance evaluation of models is typically measured using various metrics such as accuracy, precision, recall, F1 score, number of correct predictions, and number of incorrect predictions. Overall Accuracy Score is the most common accuracy measure and represents the percentage of correct predictions made by the model on the entire dataset. It gives an overall assessment of the model's performance. Precision and recall are commonly used for evaluating models in binary classification tasks. Precision represents the proportion of true positives out of all positive predictions, indicating how well the model identifies relevant instances. Recall, also known as sensitivity, calculates the proportion of true positives detected out of all actual positive instances, showing the model's ability to find all relevant cases. The F1 score combines precision and recall into a single metric, providing an overall evaluation of the model's performance. It is the harmonic mean of precision and recall, giving equal weight to both measures.

In our case, we applied two classifiers to the dataset in order to detect instances of racism. Among these classifiers, Logistic Regression achieved the highest score with an accuracy of 93 percent, while Bernoulli Naive Bayes scored slightly lower at 92 percent. Despite the minor difference in accuracy, both classifiers demonstrate good performance on the dataset. The accuracy results for both classifiers can be observed in Table 1.

TABLE I
Model Performance

Model	**Acc. Score**	**F1 Score**	**Recall**	**Precision**
Logistic Regression	93%	0.93	0.93	0.92
Bernoulli Naive Bayes	92%	0.91	0.92	0.91

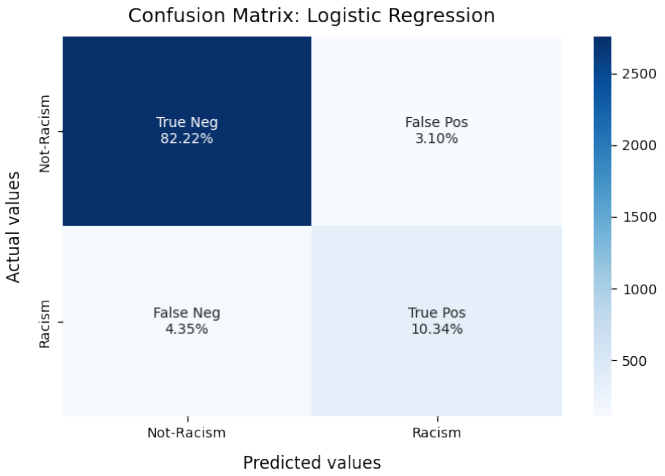

Fig. 1. Confusion Matrix of Logistic Regression

A. *Validation using Confusion Matrix*

This phase plays a pivotal role in objectively determining the performance and reliability of a model. To assess the models' performance, we applied a statistical validation approach. In order to make an empirical comparison and select the best model, we utilized the Confusion Matrix for Validation technique. This approach allowed us to gain deeper insights into the models' performance and make an informed decision.

Confusion matrix summarizes the performance of a machine learning model by providing a detailed breakdown of predictions. It displays the counts of true positives, true negatives, false positives, and false negatives, which allows for a deeper analysis of the model's performance on different classes or categories. The confusion matrix helps in identifying any patterns or imbalances in the model's prediction accuracy and providing insights into specific types of errors that the model might be making.

B. *Result Analysis with Explainable AI*

Explainable Artificial Intelligence (XAI) focuses on developing machine learning models and algorithms that can provide transparent and interpretable results and insights. By incorporating XAI techniques, users can gain a deeper understanding of how an AI system arrived at a particular decision or prediction, improving trust and accountability in AI applications.

Explainable Artificial Intelligence is employed to gain an enhanced understanding of a model's output and the underlying reasoning that leads to its conclusions. Utilizing the LIME framework for XAI, we gain insights into the model's outputs and the logic behind the specific results based on the words in the sentence. In this study, LIME was exclusively applied to

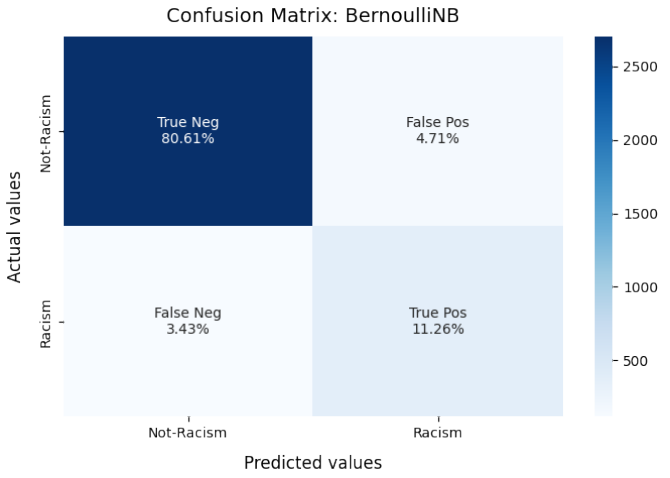

Fig. 2. Confusion Matrix of Bernoulli Naive Bayes

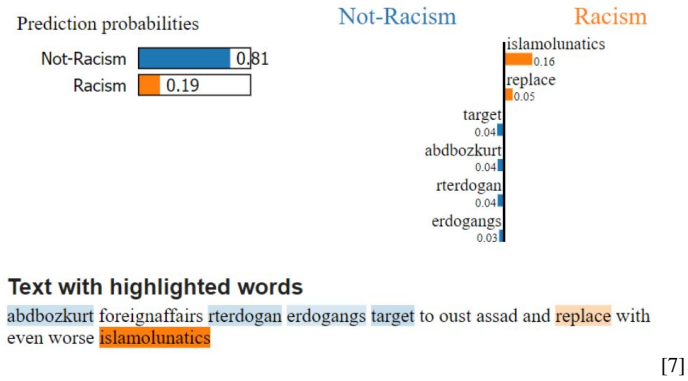

Fig. 3. LIME analysing the classifiers for a single instance(Sample No. 36)

examine the output of the logistic regression model, enabling us to provide specific examples that shed light on the model's decision-making process.

In order to enhance our understanding of the model, we leveraged XAI techniques, employing LIME specifically to interpret the output of the Logistic Regression model. LIME facilitates the explanation of individual predictions made by the machine learning model. To ensure clarity, we created two distinct labels "racism" and "not racism", clearly defining the classes. Our dataset consisted of 13,471 rows and 5 columns, with 75% of the data reserved for training. By utilizing LIME, we were able to predict the ability of a single instance and provide a comprehensible explanation for the model's decision-making process.

In Figure 3, the model predicts that this particular sentence contains words associated with non-racism with a confidence of 81%, while racist words contribute to 19% of the prediction. This prediction is further explained by showcasing the percentage contribution of specific words. On the right side, we can observe those words that play a significant role in the model's prediction.

In Figure 4, the model confidently predicts that this specific sentence contains words related to racism with a high confi-

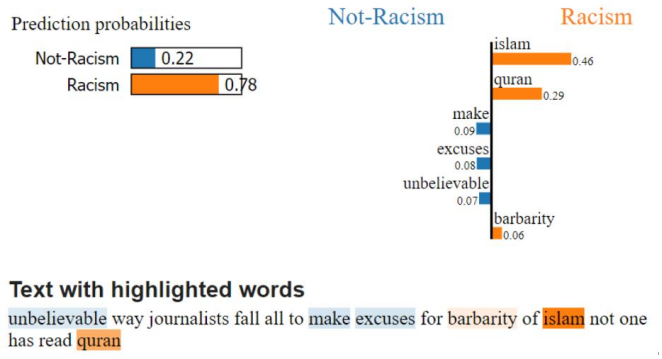

Fig. 4. LIME analysing the classifiers for a single instance(Sample No. 1876)

dence of 78%, while non-racist words contribute to just 22% of the prediction. To provide an explanation for this prediction, the model showcases the percentage contribution of specific words. On the right side, we can observe the words identified by the model that strongly influenced its prediction.

REFERENCES

[1]
[2]
[3]
[4]
[5] Pérez-Landa, G.I., Loyola-González, O., Medina-Pérez, M.A. (2021). An Explainable Artificial Intelligence Model for Detecting Xenophobic Tweets. Applied Sciences, 11(22), 10801. https://doi.org/10.3390/app11221080
[6] https://qspace.library.queensu.ca/bitstream/handle/1974/31678/Islam$_M ohammadRaf$ 3
[7]