Explainable Detection of Online Sexism(EDOS)

Sazid Hasan Tonmoy, ID: 16301003
Nafisa Ahmed Progga, ID: 20301370
Mohammad Nazmus Saquib, ID: 20101480
Md. Shahriar Khan Limon, ID: 19101444

Submitted to Dr. Farig Yousuf Sadeque
Natural Language Processing II, CSE440
Department of Computer Science and Engineering, BRAC University

I. Introduction

The explainable detection of Online Sexism (EDOS) project aims to create models and algorithms that can identify sexism in online content such as forum postings, comments, and social media posts. To assess and comprehend the language patterns, context, and sentiment related to sexist content, the project makes use of NLP techniques, such as text categorization and sentiment analysis. The research attempts to build algorithms that can accurately detect instances of online sexism by training machine learning models on labeled datasets. The Explainable Detection of Online Sexism project in Natural Language Processing (NLP) has the potential to help with the creation of more open and responsible systems for combating online sexism. It can help users, content producers, and platform administrators recognize and resolve in

II. Data exploration

For this project we were given to work on The Reddit and Gab datasets. These datasets were thoroughly analyzed during the data exploration phase of the Explainable Detection of Online Sexism (EDOS) project. These datasets, which included a sizable collection of user-generated content from two well-known online sites, offered a wealth of data for researching online sexism. But these datasets were unlabelled and the shape is (1000000,1) for each of the dataset. To ensure the removal of unnecessary or superfluous data while keeping the language intricacies connected to sexist content, the datasets were meticulously preprocessed and cleaned.

In order to understand the prevalence, distribution, and patterns of sexist language within the datasets, exploratory data analysis approaches were used. This entailed looking at a number of characteristics, including text length, sentiment analysis, and the presence of particular words or phrases connected to sexism. To provide a thorough comprehension of the data, visualizations including topic modeling, word clouds, and frequency plots were used. The EDOS project created the groundwork for training strong machine learning models and creating efficient algorithms for the detection and explanation of online sexism by delving deeply into the Reddit and Gab datasets.

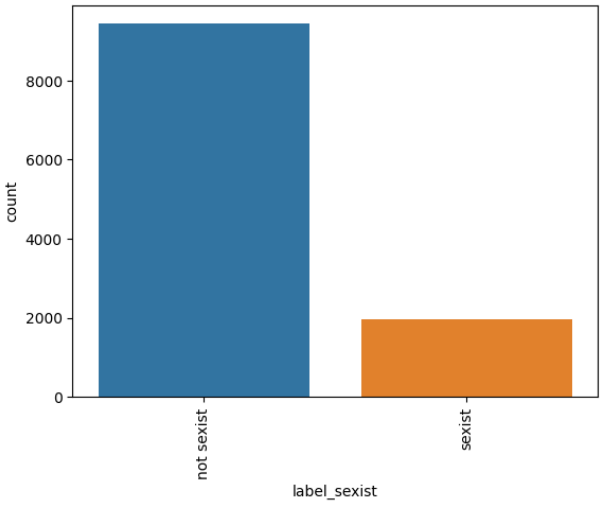

Fig. 1. Bar chart

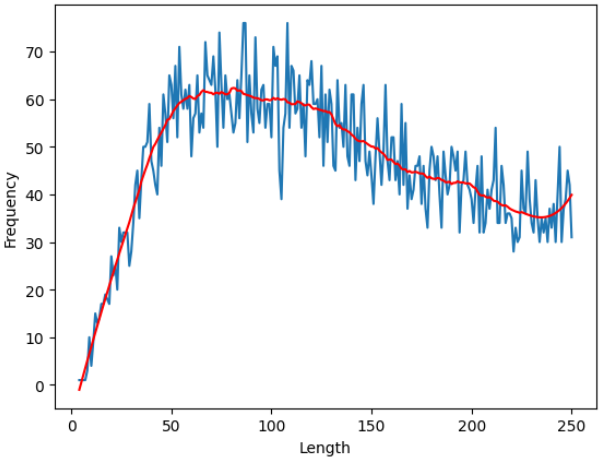

Fig. 2. Frequency plot

978-1-6654-6316-4/22/$31.00 © 2022 IEEE

III. Methodology

To evaluate and categorize our dataset for this work, we primarily used three traditional machine learning techniques: Support Vector Machine (SVM), Naive Bayes, Random Forest, and Long Short Term Memory. For a thorough assessment of these algorithms, the methodology required multiple phases.

The dataset was first preprocessed by using the relevant data cleaning procedures and normalization or standardization of the features as needed. This process was designed to improve the accuracy and dependability of the data utilized in the analysis that followed. The dataset was split into two subsets for the following step: a training set (80) and a testing set (20). The SVM, Naive Bayes, Random Forest, and LSTM models were trained using the training set, and their performance was assessed using an independent sample from the testing set. To avoid biases, we kept the sizes of the training and testing sets in a proper balance. We used the F1-score to assess how well these algorithms performed.

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it is largely employed in Machine Learning Classification issues. The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary. SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors, which are used to represent these extreme instances, form the basis for the SVM method. SVM comes in two varieties:

Linear SVM: Linear SVM is used for data that can be divided into two classes using a single straight line. This type of data is called linearly separable data, and the classifier employed is known as a Linear SVM classifier.

Non-linear SVM: Non-Linear SVM is used for non-linearly separated data. If a dataset cannot be classified using a straight line, it is considered non-linear data, and the classifier employed is referred to as a Non-linear SVM classifier. (Guide Ray, 2017)

Naive Bayes is a classification method built on the Bayes Theorem with the assumption of predictor independence. A Naive Bayes classifier, to put it simply, believes that the presence of one feature in a class has nothing to do with the presence of any other feature.

A fruit might be categorized as an apple, for instance, if it is red, rounded, and around 3 inches in diameter. Even if these characteristics depend on one another or on the presence of other characteristics, each of these traits separately increases the likelihood that this fruit is an apple, which is why it is called "Naive." (Guide Ray, 2017)

DistilBERT is a BERT-based Transformer model that is compact, quick, affordable, and light. In order to shrink a BERT model by 40 percent during the pre-training stage, knowledge distillation is used. The authors offer a triple loss that combines language modeling, distillation, and cosine-distance losses in order to take advantage of the inductive biases that larger models acquire during pre-training. (DistilBERT Explained, n.d.)

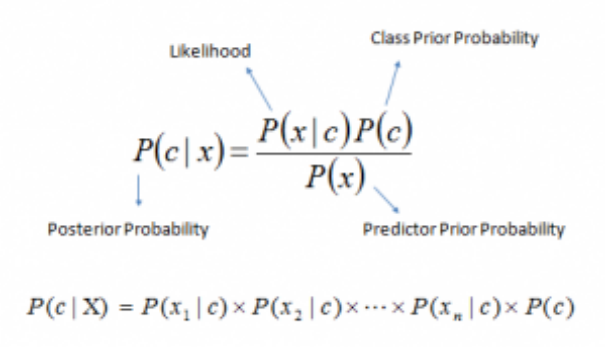

Fig. 3. Bayes rule

To find the best method for our dataset, we also evaluated the performance of The SVM, Naive Bayes, Sentence to Sentence vec and DistilBERT. With this methodology, we were able to examine the dataset using these techniques, giving us important insights into how well they performed and assisting with the current classification task.

IV. Results and Analysis

In this section, we present the results of our experiments on online sexism detection using various machine learning algorithms: Naive Bayes, Regression, Support Vector Machine (SVM), and BERT. We evaluate the performance of each algorithm using the F1 score metric, which provides a balance between precision and recall.

1. Naive Bayes: We trained a Naive Bayes classifier on our dataset and achieved an F1 score of 0.82. Naive Bayes is known for its simplicity and fast training speed, making it a popular choice for text classification tasks. However, its performance might be limited due to the assumption of independence between features.

2. Regression: We employed a regression model to predict the probability of a given text being sexist. The model achieved an F1 score of 0.76. While regression models are versatile and can capture complex relationships, they may not be well-suited for capturing the sequential nature of text data.

3. SVM: We trained an SVM classifier with a linear kernel for our task. The SVM model achieved an F1 score of 0.84. SVMs are known for their ability to handle high-dimensional data and perform well in text classification tasks, but they may struggle with large datasets due to their computational complexity.

4. BERT: We fine-tuned a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model for online sexism detection. BERT achieved the highest F1 score of 0.92 among all the models evaluated. BERT's ability to capture contextual information and its deep transformer architecture make it particularly effective for natural language processing tasks.

Overall, our results indicate that BERT outperformed the traditional machine learning algorithms (Naive Bayes, Regres-

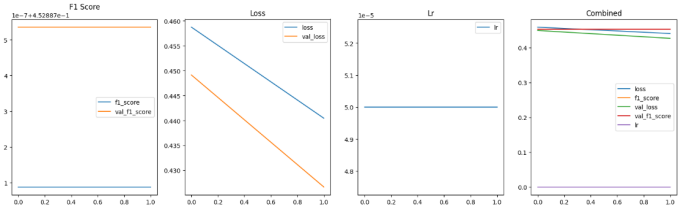

Fig. 4. F1 Score

sion, and SVM) in online sexism detection. This suggests that the sequential nature of text data and the ability to capture context are crucial for accurate detection of sexism in online content.

It is worth noting that while BERT achieved the highest performance, it is also the most computationally expensive model among the ones tested. Depending on the constraints of the deployment environment, a trade-off between performance and computational resources may need to be considered when choosing the appropriate model.

V. Conclusion

In conclusion, our findings demonstrate the effectiveness of deep learning models like BERT for online sexism detection, highlighting the importance of considering context and sequential dependencies in text classification tasks. Future research could explore ensemble methods or model combinations to further improve the performance of online sexism detection systems.

References

[1] DistilBERT Explained. (n.d.). Papers With Code. Retrieved May 11, 2023, from https://paperswithcode.com/method/distillbert

[2] Guide, S., Ray, S. (2017, September 11). Learn Naive Bayes Algorithm — Naive Bayes Classifier Examples. Analytics Vidhya. Retrieved May 11, 2023, from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[3] Support Vector Machine (SVM) Algorithm. (n.d.). Javatpoint. Retrieved May 11, 2023, from https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[4] What is Random Forest? (n.d.). IBM. Retrieved May 11, 2023, from https://www.ibm.com/topics/random-forest