Sexism Text Classification

Albert Boateng - 20216003

May, 2023

1 Introduction

Sexism refers to discrimination or prejudice based on someone's gender or sex. Many authors consider it a type of hate speech because it aims to dehumanize and marginalize a group of people based on their gender. This type of language and behavior can be incredibly harmful, leading to negative self-esteem, mental health issues, and a decreased sense of worth or value. As a result, detecting and classifying sexist posts has become an important research topic to improve the quality of social media interactions.

Detecting sexism in social media posts is a challenging task due to various factors such as the presence of multiple labels, the use of sarcasm, irony, abbreviations, emojis, misspellings, and memes. These factors often make it difficult to classify posts accurately into different classes. The complexity of the task makes it a crucial research topic. Therefore, many models, techniques, and methodologies have been proposed to address this issue. One of the popular approaches is to use machine learning algorithms, which can learn patterns from the data and classify the posts accurately. Natural language processing techniques are also used to extract features from text data and build models based on those features. Furthermore, deep learning approaches such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been applied to social media text data to improve the accuracy of classification. In our study, machine learning models such as Logistic Regression, Naive Bayes, Support Vector Machine (SVM), and AdaBoost Classifiers will be used for our classification.

2 Datasets

The dataset used for the classification is the SemEval 2023 - Task 10 - Explainable Detection of Online Sexism dataset. It contains 14,000 observations and 5 features ('rewire_id', 'text', 'label_sexist', 'label_category', 'label_vector'). The 'text' column contains 14,000 sexist/non-sexist text. Out of the 14,000 values in the target variable ('label_sexist'), 10,602 are 'not sexist' and 3,398 are 'sexist'. Also, in the target variable ('label_category'), 10,602 are 'none' labelled category, 310 are '1. threats, plans to harm and incitement', 1,590 are '2. derogation', 1,165 are '3. animosity' and 333 are '4. prejudiced discussions'.

For the target variable ('label_vector'), 10,602 are 'none' labelled vector, 56 are '1.1 threats of harm', 254 are '1.2 incitement and encouragement of harm', 717 are '2.1 descriptive attacks', 673 are '2.2 aggressive and emotive attacks', 200 are '2.3 dehumanising attacks & overt sexual objectification', 637 are '3.1 casual use of gendered slurs, profanities, and insults', 417 are '3.2 immutable gender differences and gender stereotypes', 64 are '3.3 backhanded gendered compliments', 47 are '3.4 condescending explanations or unwelcome advice', 75 are '4.1 supporting mistreatment of individual women', and 258 are '4.2 supporting systemic discrimination against women as a group'. For simplicity sake, we will be using the targets ('text', 'label_sexist') for our analysis.

3 Methodology

The columns 'text' and 'label_sexist' were first selected from the dataset for analysis. The sentences in the 'text' were then preprocessed. This was done by removing special characters in the sentences, lower-casing all the words in the sentences, tokenizing the sentences into words, as well as removing all the stop-words in the sentences. The results obtained were then lemmatized to reduce the words to their root form. The 'label_sexist' column was also converted into Numeric values by mapping 'not sexist' values to 0's and 'sexist' values to 1's. Because the dataset isn't balanced (10,602 'not sexist' against 3,398 'sexist'), minority over-sampling balancing technique was applied to the dataset to oversample the minority class to balance the dataset.

After these, our dataset was spiltted into training and testing set in the ratio 0.8 to 0.2 respectively. The training set was then further divided into training and development set in the ratio 0.9 to 0.1 respectively. The X training, X development, and X testing sets were then converted into numeric features using the TfidfVectorizer which is a combination of CountVectorizer and Tf-idf transformer. The CountVectorizer is the bag of word model which convert the text data into a count matrix while the Tf-idf transformer transform the count matrix to a normalized tf or tf-idf representation. The results obtained were then passed through four Machine Learning models: Logistic Regression, Naive Bayes (Multinomial Naive Bayes), Support Vector Machine (SVM, kernel='linear'), and AdaBoost Classifier (n_estimators=50 and learning_rate=1) to compare and contrast their performance.

4 Results

For Logistic Regression: Of the 4,241 testing data points, 1,767 of them were correctly classified as Non Sexist (True Negative), 1,821 were correctly classified Sexist (True Positive), 362 were wrongly classified Sexist (False Positive), and 291 were wrongly classified Non Sexist (False Negative) as shown in the Confusion Matrix below (Figure 1). This leads to an Accuracy of 85% and a Macro F1-Score of 85%.

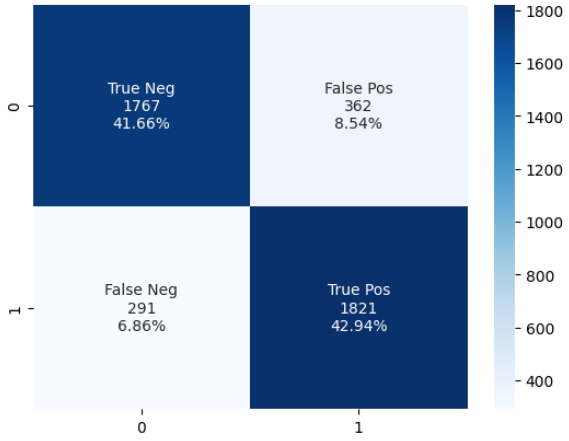

Figure 1: Confusion Matrix of Logistic Regression

For Naive Bayes classifier: Of the 4,241 testing data points, 1,695 of them were correctly classified as Non Sexist (True Negative), 1,531 were correctly classified Sexist (True Positive), 434 were wrongly classified Sexist (False Positive), and 581 were wrongly classified Non Sexist (False Negative) as shown in Figure 2. This leads to an Accuracy of 79% and a Macro F1-Score of 78%.

For AdaBoost Classifier: Of the 4,241 testing data points, 1,695 of them were correctly classified as Non Sexist (True Negative), 1,531 were correctly classified Sexist (True Pos-

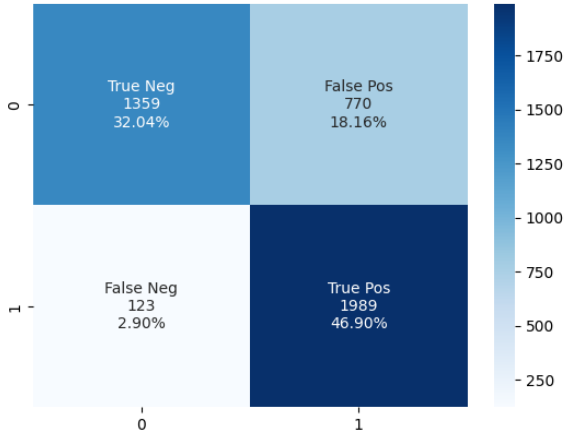

Figure 2: Confusion Matrix of Naive Bayes classifier

itive), 434 were wrongly classified Sexist (False Positive), and 581 were wrongly classified Non Sexist (False Negative) as shown in Figure 3. This leads to an Accuracy of 76% and a Macro F1-Score of 76%.

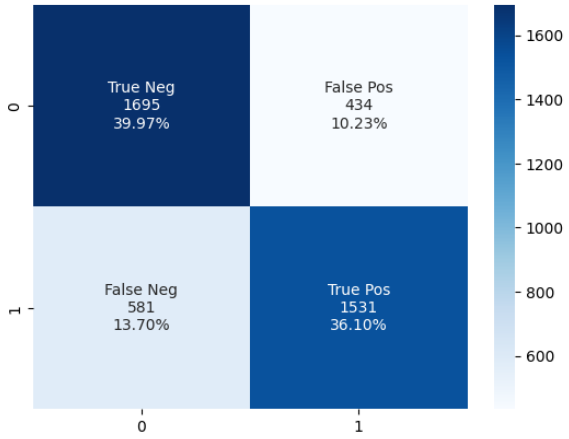

Figure 3: Confusion Matrix of AdaBoost Classifier

For Support Vector Machine (SVM): Of the 4,241 testing data points, 1,751 of them were correctly classified as Non Sexist (True Negative), 1,911 were correctly classified Sexist (True Positive), 378 were wrongly classified Sexist (False Positive), and 201 were wrongly classified Non Sexist (False Negative) as shown in the Confusion Matrix below (Figure 2). This leads to an Accuracy of 86% and a Macro F1-Score of 86%.

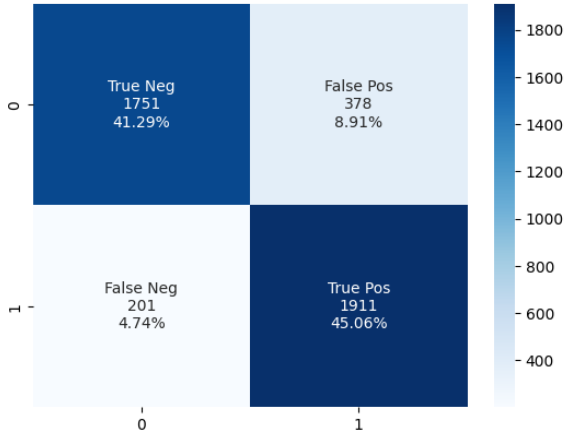

Figure 4: Confusion Matrix of Support Vector Machine (SVM)

Comparing the performance of the algorithms shows that SVM performed best followed by Logistic Regression, then Naive Bayes, then finally AdaBoost Classifier as shown in figure 5. Also, it is important to note that most of the sampling done were random and as such these values might change slightly running the codes again.

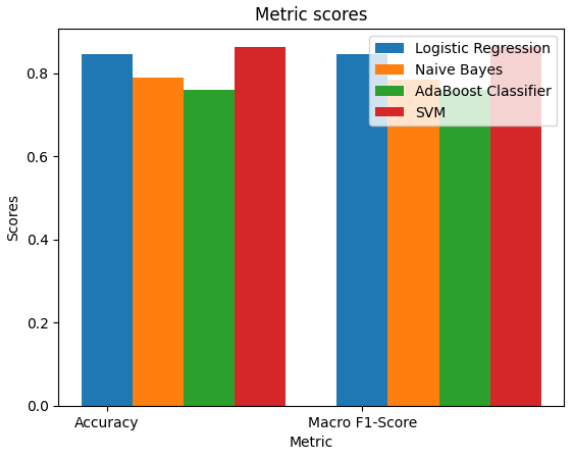

Figure 5: Bar Graph Comparing the Performance of the Models

5 Reference

- https://codalab.lisn.upsaclay.fr/competitions/7124
- https://www.datacamp.com/tutorial/text-classification-python
- https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a
- https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/
- https://stackabuse.com/text-classification-with-python-and-scikit-learn/
- https://arxiv.org/pdf/2303.04222.pdf