Data Cleaning + Exploration

Perform data cleaning and exploration by checking for null values and duplicate texts within the dataset. Check the distributions for the different types of cyberbullying and plot them using a pie chart. The distributions were roughly equal, with each type accounting for around 17%, except for "other_cyberbullying," which accounted for 13.6%. The tweets in the dataset were stripped of punctuation, pictures, and HTML links. They were also tokenized, stopwords were removed, and their part of speech was identified. Finally, the words were lemmatized, and their sentiment scores were analyzed. The sentiment distribution shows a unimodal pattern with a slight negative tail. This is expected since the dataset consists of 1/6 not_cyberbullying and 5/6 cyberbullying instances. Word clouds were created for each type, highlighting the most common words. For example, for non_cyberbullying, common words included "mkr," "bully," "like," "get," "school," and "people," which were neutral words.

Multi-Classification

1. **Naive Bayes (MultinomialNB)**

   The data was processed using LabelEncoder and split into training and test sets. The lemmatized data was then converted into numerical features using TfidfVectorizer for both the X_train and X_test datasets. Hyperparameter tuning and cross-validation were performed on the Naive Bayes model based on MultinomialNB(). The best model and parameters were discovered, and predictions were made using this model. To evaluate the performance of the model, the classification report was utilized, which provided an accuracy score of 0.75. Although this accuracy score was decent, it did not meet my expectations. To visualize the model, a probability distribution was created. It revealed that the top three features with the highest probabilities were the 4th, 3rd, and 1st/5th, which tied for 3rd place.

2. **Random Forest**

Next, the Random Forest model was created using the same data as the Naive Bayes model. The Random Forest model was evaluated using the classification report, which showed an accuracy of 0.84. This accuracy is considered decent and meets your expectations. After evaluating the model, the feature importances were obtained and stored in a new dataframe. This dataframe will be used for plotting. The top 50 important words in the Random Forest model were identified, which included words like “school”, “nigg**”, “bully”, “dumb”, “high”, “muslim”, “gay”, “girl” and many more.

3. **Gradient Boosting**

   The Gradient Boosting model was created using a similar process as the Random Forest model. The classification report shows that the Gradient Boosting model performed slightly worse than the Random Forest model, achieving an accuracy score of 0.83. However, it is still considered an amazing result. Similar to the Random Forest model, the feature importances for the Gradient Boosting model were discovered. The top 50 words used in the model were plotted, and they shared some similarities with the Random Forest model. The top word for both models was "school," followed by "nigg**" and included other words from the Random Forest model but in a different order. For example, after "school," the top words for the Gradient Boosting model were "nigg**," "rape," "muslim," "christian," "mkr," "female," "bully," "sexist," and many more.

Binary Classification

1. **SVM**

   The data was preprocessed by creating a function to split it into 0s and 1s based on whether it is part of the cyberbullying type. This function was then applied to the dataframe, creating a new column called "binary_encode". The data was then split into training and testing sets and converted to numerical values using TfidfVectorizer. The SVM model was created using this preprocessed data, and its performance was evaluated using a classification report. The report

shows that the accuracy of the SVM model was 0.87, which is considered amazing. To further analyze the performance of the model, a confusion matrix was plotted. The confusion matrix shows that the predictions were largely accurate. It predicted the 1s accurately for 7389 out of 9204 instances, and the 0s accurately for 616 out of 9204 instances. Combined, the model correctly predicted 8005 out of 9204 instances. This indicates that the SVM model performed well in predicting the binary labels, achieving a high accuracy and making accurate predictions for both cyberbullying (1s) and non-cyberbullying (0s) instances. The precision-recall curve also shows a high precision overall of 0.825.

2. **Logistic Regression**

   Similar procedures were applied for the Logistic Regression model as well. The classification report revealed that the accuracy of the Logistic Regression model was 0.86, which was slightly lower than the SVM model but still considered good. To further analyze the performance of the model, a precision-recall curve was plotted and compared with the SVM model's curve. The two curves showed subtle differences, indicating that their precision and recall values were roughly similar. The precision-recall curve is a useful tool for evaluating the trade-off between precision and recall for different classification thresholds. It helps us understand how the model's performance changes as we adjust the threshold for classifying instances. Overall, the Logistic Regression model achieved a good accuracy score and showed comparable precision and recall to the SVM model, indicating its effectiveness in predicting the binary labels.