Major Project

Project Name:

Machine Learning May Major Project

Problem Statement:

Create a classification model to predict the type of cyberbullying (age, ethnicity, gender, religion, not cyberbullying, other cyberbullying)

Context: As social media usage becomes increasingly prevalent in every age group; a vast majority of citizens rely on this essential medium for day-to-day communication. Social media's ubiquity means that cyberbullying can effectively impact anyone at any time or anywhere, and the relative anonymity of the internet makes such personal attacks more difficult to stop than traditional bullying.
On April 15th, 2020, UNICEF issued a warning in response to the increased risk of cyberbullying during the COVID-19 pandemic due to widespread school closures, increased screen time, and decreased face-to-face social interaction. The statistics of cyberbullying are outright alarming: 36.5% of middle and high school students have felt cyberbullied and 87% have observed cyberbullying, with effects ranging from decreased academic performance to depression to suicidal thoughts.

Dataset:

https://drive.google.com/file/d/1Igs94JugaCyApLezkysQAl5Sz13G_8FC/view?usp=sharing

Details of features:

The columns are described as follows:

1. tweets: User Tweets
2. cyber bullying - type: age, ethnicity, gender, religion, not cyberbullying, other cyberbullying

Steps to consider:

1. Read the dataset
2. Remove handle null values (if any).
3. Preprocess the disaster tweets data based on the following parameter: a) Tokenizing words
   b) Convert words to lower case
   c) Removing Punctuations
   d) Removing Stop words
   e) Stemming or lemmatizing the words
4. Transform the words into vectors using
   a) Count Vectorizer
   OR
   b) TF-IDF Vectorizer
5. Select x (independent feature) as tweets after preprocessing and cyberbullying type as y (dependent feature).
6. Split data into training and test data.

7. Apply the following models on the training dataset and generate the predicted value for the test dataset

   a) Multinomial Naïve Bayes Classification

   b) Logistic Regression

   c) Decision Tree Classification

   d) Random Forest Classification

8. Predict the cyberbullying type for test data

9. Compute Confusion matrix and classification report for each of these models

10. Report the model with the best accuracy.