Amritha Prakash

Project Proposal: Application for determining the toxicity of a comment. This indirectly would allow the users to flag those comments as bad or even report them if used in an inappropriate manner.

Aim: The idea behind this project is to predict if the comment added on the application is toxic, or not? If yes, then understand which category of toxicity it falls under. Also, we can try to figure out the percentage of toxicity.

Steps that would be followed -

1. Data Cleaning
2. Explore the data
3. Make use of the dataset and try to train the model for all the different ML classifiers. i.e. Choose a Model between the different Machine Learning classifiers based on the accuracy that they provide like for Logistic Regression, KNN, SVM, Naïve Bayes, etc
   Otherwise make use of pre-trained model, to train the model for better accuracy.
4. Based on accuracy for each model, build and deploy the model with an accurate one to be utilized within an application.
5. Hence, further create an application to learn to categorize on that same aspect if a new comment has been put in the system.

Reason for creating the application:
Social media's widespread use has made it possible for people to freely share their thoughts online. But concurrently, this has led to the rise of hate and conflict, which has made online spaces unwelcoming to users. Even though hate crimes occur on a variety of platforms, there aren't many models available for detecting hate online.

Online hatred has been identified as a significant problem on online social media platforms. It is characterized by abusive language, hostility, cyberbullying, hatefulness, and many other things. Social media platforms are the primary venues for this kind of harmful conduct.
Toxicity is defined as anything rude, disrespectful or otherwise likely to make someone leave a discussion. The significance of creating the application is to provide a way to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.

Constraints:

- Availability of Data: It can be difficult to get high-quality labeled data for toxicity detection model training. It can take a while to create a comprehensive dataset that addresses several types of toxicity (such as hate speech, foul language, and cyberbullying), and it might need to be manually annotated.
- It can be arbitrary and culturally specific to define what is poisonous. It is important to make sure the training data and algorithms of the model don't unintentionally reinforce prejudices or stifle free speech.

- It can be challenging to achieve high accuracy in toxicity detection, particularly for lower forms of toxicity or false positives/negatives. Model performance is greatly influenced by the selection of machine learning models and the caliber of features.
- It is crucial to make sure that the model does not reveal bias in its predictions or unfairly target groups. An obstacle in the development of AI is addressing issues of fairness and bias.

Goals :

- To make the internet a safer place, build a model that can reliably identify harmful remarks, such as hate speech, derogatory language, and cyberbullying.
- Toxic comments can be categorized into many levels of toxicity, allowing for a more nuanced response.
- Reduce false positives and false negatives by recognizing poisonous comments with a high degree of accuracy.
- To promote user participation and effective reporting, design an application with a simple and user-friendly interface. So that,
- Reduce bias and make sure that the toxicity detection model handles every user equally—that is, without singling out demographics.

Non - goals :

- While accuracy is the goal, it might not be possible to do at 100% accuracy. Decide on a reasonable cutoff point for acceptable accuracy.
- The project has no intention of suppressing dissenting views or lawful free speech. It ought to focus solely on offensive material.
- The application's main objective is to offer insights into whether a comment is toxic, not to make decisions for users if it meant the same.
- Getting 100% of users to accept or agree with the application's toxicity assessments shouldn't be the aim. Divergent views are normal, but the system needs to strive for fairness and reason.
- Refrain from introducing extraneous complexity or features that don't support the application's primary objectives.

Documentation for Dataset:

The data set contains the training set, which has approximately 1,59,000 samples and the test set which contains nearly 1,53,000samples. All the data samples contain 8 fields which include ‘Id’, ‘Comments’, ‘Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’ and ‘Loathe’.

Columns -

- Malignant: It is the Label column, which includes values 0 and 1, denoting if the comment is malignant or not.
- Highly Malignant: It denotes comments that are highly malignant and hurtful.

- Rude: It denotes comments that are very rude and offensive.
- Threat: It contains indication of the comments that are giving any threat to someone.
- Abuse: It is for comments that are abusive in nature.
- Loathe: It describes the comments which are hateful and loathing in nature.
- ID: It includes unique Ids associated with each comment text given.
- Comment text: This column contains the comments extracted from various social media platforms.

Reference of Dataset -

(2020). Dataset [Data set]. Kaggle.
https://www.kaggle.com/datasets/surekharamireddy/malignant-comment-classification/data