Project 2:

The prevalence of online toxicity has become a growing concern in social media platforms, and it can negatively impact individuals and communities. In this assignment, you will work with a toxic comment classification dataset that includes comments labeled with multiple toxic classes, and use Natural Language Processing (NLP) techniques to build a multilabel predictive model.

The dataset used in this assignment contains a large number of comments from various online platforms, such as Wikipedia, Reddit, and Twitter. Each comment is labeled with multiple classes of toxicity, including severe toxicity, obscene, threat, insult, identity hate, and others. A comment may belong to one or more toxicity classes, making this a multilabel classification task.

The main goal of this assignment is to develop a predictive model that can accurately classify toxic comments into one or more toxicity classes in the dataset, using some innovative methods. To achieve this goal, you need to follow the following steps:

1. Data Exploration and Preparation:

- Explore the dataset and analyze the distribution of each toxicity class.
- Preprocess the text data by removing stopwords, punctuation and other irrelevant characters. Also, perform text normalization techniques such as stemming and lemmatization and can also use other preprocessing techniques like De-contraction, Chunking, etc.
- Split the data into training and testing sets.

2. Feature Extraction:

- Extract relevant features from the preprocessed text data, such as Bag-of-Words, TF-IDF, and word embeddings.

3. Model Building:

- Build various models, using Multi-label Classification Techniques like, Problem Transformation methods, ensemble algorithms, Adapted algorithms, Neural Networks.
- Experiment with innovative models, such as neural networks, deep learning models, or ensemble methods, and compare their performance with traditional models.
- Evaluate the performance of each model. Fine-tune the best performing model.

4. Comparing Models and concluding:

- Compare the best model and predict their toxicity classes.,
- Build a flask Api.

Dataset:

The Jigsaw Toxic Comment Classification Challenge dataset contains over 1.6 million comments from Wikipedia talk pages. The dataset is labeled with types of toxicity:

i. toxic
ii. severe_toxic
iii. obscene
iv. threat
v. insult
vi. identity_hate

Data link: Toxic Comment Classification