Project name: Toxic Comment Classification

Participants:

Lavanya:

Manasa:

Manumitha Kamarapu : manumithakamarapu@my.unt.edu

**Project Abstract**

The goal of the toxic comment classification project is to develop a system that can identify and flag toxic comments on online platforms. The project aims to improve the user experience and create a safe and respectful environment for all users. The system will identify comments that contain hate speech, harassment, and other forms of toxic language, allowing moderators to quickly review and remove them.

To achieve this goal, the project will involve several key actions. First, we will gather a large dataset of comments using Kaggle platform, which includes comments from various social media, discussion forums, and blogs. These comments will be labeled as toxic or non-toxic. Then, we will train machine learning models using Natural language processing techniques to classify comments as toxic or non-toxic. Finally, we will test the different models and come to conclusion which model is better for the classification and gives better accuracy.

The workflow diagram for the toxic comment classification project is shown below.

Evaluation and final results

Classification using various models

Testing Data

Training Data

Data collection (i.e., raw comments )

Feature Extraction (tokenization, stop words, etc.)

Data processing

Overall, the toxic comment classification project aims to create a safer and more respectful online environment for all users. By identifying and flagging toxic comments, the system will help reduce harassment, hate speech, and other forms of toxic language on online platforms.

**Data Abstraction**

The Jigsaw Unintended Bias in Toxicity Classification dataset is a publicly available dataset consisting of approximately 1.8 million comments from the online discussion platform, Civil Comments. The dataset was created by Jigsaw, a subsidiary of Alphabet Inc. (formerly known as Google Ideas).

Types of dataset:

The dataset includes both the comments and associated toxicity scores, which were assigned by human annotators. The toxicity scores range from 0 to 1, with 0 indicating a non-toxic comment and 1 indicating a highly toxic comment. Additionally, the dataset includes demographic information about the commenters, such as their gender, age, and education level.

There are two types of toxicity labels in the dataset: binary and continuous. The binary toxicity label indicates whether a comment is toxic or non-toxic, while the continuous toxicity label is a real-valued number between 0 and 1, indicating the degree of toxicity of a comment. The dataset also includes several additional attributes, including the comment text, the date the comment was made, and the comment ID.

Attributes of dataset:

One important attribute of the Jigsaw dataset is the presence of unintended bias. The dataset was created by crowdsourcing the annotation task to a diverse group of workers, which can lead to bias in the annotations. Additionally, the dataset includes demographic information about the commenters, which can be used to study the relationship between demographic factors and toxic behavior. This has implications for the ethical use of the dataset and the development of machine learning models that use it.

**Project Design**

Programming Logic:

The project design for the toxic comment classification project can involve several machine learning models to achieve the goal of identifying and flagging toxic comments. Below is an overview of a potential project design that incorporates different machine learning models:

Data Collection: Collect comments from Kaggle platform, which includes comments from various social media, discussion forums, and blogs.

Data Preprocessing: Preprocess the comments to clean the data and remove irrelevant information. This involves techniques such as tokenization, stemming, removing stop words, and normalization.

Feature Extraction: Extract features from the preprocessed data that can be used to train the machine learning models. This involves techniques such as bag-of-words, TFIDF, and LSTM.

Model Training: Train different machine learning models using the preprocessed data and extracted features. Some potential models that we consider include:

Logistic Regression: A linear model that is commonly used in text classification tasks.

Support Vector Machines (SVMs): A model that can separate data into different classes based on hyperplanes.

Naive Bayes: A probabilistic model that can be trained quickly and is often used in text classification tasks.

Convolutional Neural Networks (CNNs): A deep learning model that can learn hierarchical features from text data.

Recurrent Neural Networks (RNNs): A deep learning model that can capture sequential information in text data.

Model Evaluation: Evaluate the performance of the different machine learning models using metrics such as accuracy, precision, recall, and F1 score. Finally, select the best performing model for testing.

Results: Test the selected machine learning model(s) to identify and flag toxic comments on testing dataset. Hence, classify the toxic comments from all the users’ comments using the model that gives better accuracy.

Software used: Python

Packages used: pandas, numpy, sklearn, wordcloud collections, nltk and more

Modules used: stopwords, TfidfVectorizer, LogisticRegression, LinearSVC, RandomForestClassifier, DecisionTreeClassifier, WordCloud, and more

IDE used: Jupyter Notebook

**Project Milestones**

Bias in dataset:Dataset is in large size, has diversity in commenters and comments, and availability of demographic information this makes feature engineering difficult.

Word Embeddings: Due to presence of various kinds of data, vectorizing dataset is difficult.

Out-of-vocabulary words: Occurrence of words that are not defined in dictionary such as misspellings or intentionally obfuscated comments.