Group name: Solitude Ensemble

Group member’s details

Name:	Parthkumar Sutariya
Email ID:	parthsutariya31@gmail.com
Country:	India
College:	BSBI, Berlin, Germany
Specialization:	Data Analyst

Problem Description:

In the initial stages of the project, it became evident that the raw data obtained from Twitter contains various quality issues that must be addressed before further analysis. These issues include:

- Presence of missing values (NA values) in certain fields, such as tweet text or metadata.
- Inconsistencies and noise in the text data, including special characters, emojis, and irrelevant symbols.
- Potential presence of outliers or anomalies in the data distribution, which could affect model performance if not properly handled.

Approaches for Data Cleansing and Transformation:

Handling Missing Values: Various approaches have been considered to address missing values in the dataset, including:

- Mean, median, or mode imputation: Filling missing values in numerical fields with the respective statistical measures.
- Model-based imputation: Utilizing machine learning models to predict missing values based on other features in the dataset.

Text Data Cleaning: Techniques have been employed to clean and preprocess the text data, such as:

- Removal of special characters, punctuation, and non-alphanumeric symbols using regular expressions.
- Lowercasing all text to standardize the representation of words and reduce vocabulary size.
- Removing stop words and performing stemming or lemmatization to normalize text and reduce feature dimensionality.

NLP Featurization and Data Cleaning:

Featurization Techniques: Various featurization techniques have been experimented with to represent text data effectively for hate speech detection, including:

- Bag-of-Words (BoW): Representing text data as a matrix of word occurrences or frequencies.
- TF-IDF (Term Frequency-Inverse Document Frequency): Assigning weights to words based on their frequency in a document and across the entire corpus.
- Word Embeddings: Capturing semantic meanings of words by mapping them to dense vectors in a continuous space using techniques like Word2Vec or GloVe.

Data Cleaning with Regex and Python: Text data has been cleaned using regular expressions (regex) and Python programming, involving tasks such as:

- Removing special characters, punctuation, and non-alphanumeric characters.
- Lowercasing all text to standardize the representation of words.
- Removing stop words and performing stemming or lemmatization to reduce inflectional forms of words to their base or root forms.