**SENTIMENTAL ANALYSIS ON IMDB REVIEW DATA**

**Dataset:**

The dataset has two folders with 25,000 highly polar movie reviews for training, and 25,000 for testing. Both the train and test set have positive, negative samples. Also, the dataset has additional unlabeled data for use as well.

**Data Pre-Processing**

The first part before analyzing the data, requires cleaning and processing data. Cleaning data means removing noise, extra tags, augmenting and normalizing and processing data will be representing data in vector, which can be obtained by performing the necessary pre-processing tasks. In Natural Language processing (NLP), pre-processing includes techniques like Stemming and Lemmatization. Performing them in a certain order is crucial. In the given COCO dataset, the data is processed using the following steps.

1. Libraries
2. Loading the data frame
3. Lower casting
4. Expand the contractions.
5. Noise removal
6. Diacritics removal
7. Typos correction
8. Stopword removal
9. Lemmatization

**Libraries**

Initially the required libraries namely pandas, NumPy, nltk, are imported. Natural Language Toolkit (NLTK) has functions to perform tokenization, stopwords removal, stemming and lemmatization.

**Loading the data frame**

The train COCO dataset is loaded using and categorized based the sentiments i.e., whether they are positive and negative and for each data point we had rec\_id, text, sentiment, and rating as attributes.

**Lower casting**

In the lower casting phase, all the words in the data are converted into lowercase characters. For example, the line “Welcome to UTA” is converted into “welcome to uta”.

**Expand the contractions.**

The Expanding contractions, all the short forms are converted to their actual representation, for example “hadn’t” is converted into “had not”.

**Noise removal**

The phase noise removal includes multiple steps namely removing html tags, removing numbers, replacing dots with spaces, removing punctuations, and replacing duplicate spaces.

**Diacritics removal**

The nearest character is chosen and replaced by diacritics. If no character is found, they are replaced with spaces.

**Typos correction**

All the spelling mistakes are corrected, this step is crucial as stemming and lemmatization requires removing typos.

**Stopword removal**

Initially the stopwords in English are loaded from nltk library and custom stopwords are specified if necessary. Then in any given sentence, the stopwords are removed.

**Lemmatization**

Compared to stemming lemmatization is more accurate and uses dictionary-based technique, however, lemmatization is a slow process. In case of stemming, Porter stemming could have been used, however, the lemmatization is used which suits well for this dataset.

**Data Analysis:**

Once data is processed, the data is analyzed by visualizing the train data as shown in figure 1. The test data is displayed as shown in figure 2 and the classes in test data are balanced as shown in the figure 2.

A close-up of words

Description automatically generated

Fig 1. Train data Visualization

A close-up of words

Description automatically generated

Fig 2. Test data Visualization

The test data is balanced as the number of positive and negative samples are nearly for each rating from 1 to 10 as shown in the figure 3.

A graph of a test data

Description automatically generated

Fig 3. Test data is balanced.

For the train data Term Frequency-Inverse Document Frequency (TF-IDF) vectors are generated, where Term frequency (TF) measures how frequently a term word occurs in a document and Inverse Document Frequency (IDF) measures how important a term is across the entire corpus. In the next step we would explore word2vec models and would like to do more literature surveys on different Machine learning models for sentiment analysis.