Sentimental Analysis on Open-Ended Conversation

Open Ended Conversation includes surveys, reviews and questions whose answers are not specific.

Dataset: Sentiment Analysis of IMDB Movie Reviews

❖ *Introduction to Dataset*

IMDB dataset having 50K movie reviews for natural language processing or Text analytics.
This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. It provides a set of 25,000 highly polar movie reviews for training and 25,000 for testing. So, predict the number of positive and negative reviews using either classification or deep learning algorithms.

i) The Dataset consists of two columns:
   a) Review: (str) Long Review given by audience
   b) Sentiment: (str) Positive / Negative

review	sentiment
49582 unique values	2 unique values
One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. The...	positive
A wonderful little production. The filming technique is very unassuming- very old-time-B...	positive
I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air	positive

❖ ***Approach***

a) Data Engineering:

1. Cleaning Data of **stop-words** and changing to **lower case** character.
2. **Tokenize** sentences to words.
3. **Stemming** and **Lemmatization** for changing the tenses.

b) EDA using **N-gram**:

We can get some **useful information** about the words by using N-gram technique. Usually, Bi gram is preferred.

Note: Longer the context window harder it is to pick meanings.

c) Find **common words** in sentiment analysis and **form word cloud**.

d) **Word Embeddings**:

It can give **Semantic meaning** of a word.
Since every word in an open-ended conversation has an importance, word embedding approach can provide a **better relatedness** of the word with a topic.

e) Splitting Data into Train and Test

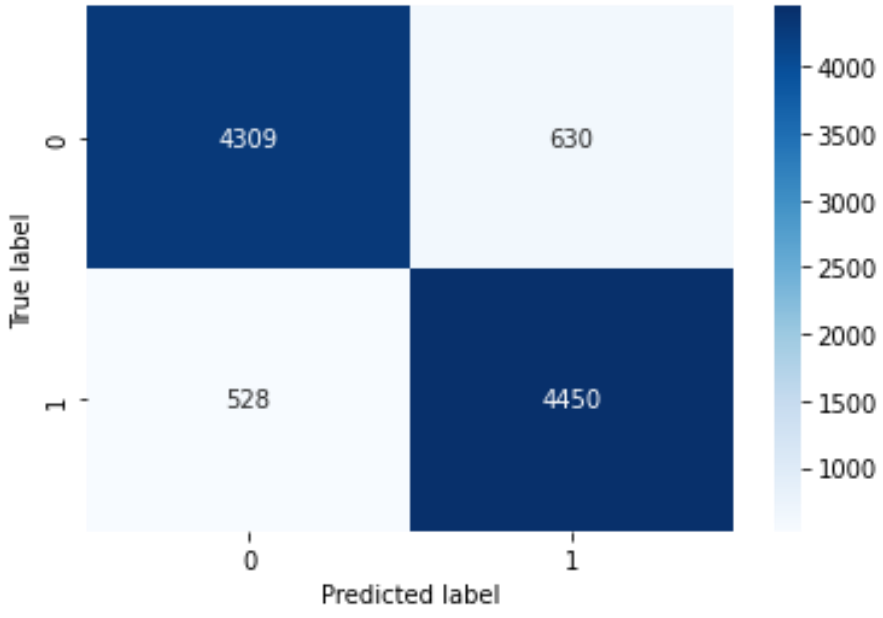
❖ ***Model Building:***

➢ **LSTM** type:

- **INPUT Layer**: Shape = Max. length of input.
- **EMBEDDING**: Perform word embedding with Vocabulary size = V+1, Output = 5, Input length = Max. length of input.
- **Batch Norm**: Performed on the layers to normalize the weighted sum of every neuron.
- **DropOut**: To Spread out weights on next layer.
- **Conv1D**: Perform convolution over one direction with stride 1 with 'RELU' activation function.
- **DropOut**: To Spread out weights on next layer.
- **Max Polling**: To extract dominating features.
- **LSTM**: Implementing LSTM by taking output dimension as 128.
- **LSTM**: Implementing LSTM by taking output dimension as 64.
- **DropOut**: To Spread out weights on next layer.
- **Dense**: Single output neuron layer with Sigmoid activation.

❖ ***Accuracy:***

The accuracy of test results was found to be 88.32%. The analysis of the result is mentioned below:



Dataset link:
https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews