NLP Report

Omar Elsayed

20031826

The objective of this natural language processing project is to build a model intended to read customer reviews for a restaurant. The goal of this modes is to be able to tell the difference between positive and negative reviews.

The Imports and key words

A screen shot of a computer

Description automatically generated

I started off the project with importing some crucial libraries such as the ‘nltk’ library that is used for NLP tasks, ‘sklearn’ and ‘tensorflow’ which both provide machine learning tools. These libraries are needed for the preprocessing and actually building and evaluating the model.

Loading the dataset

A screen shot of a computer

Description automatically generated

The next thing I did was load in the dataset that I used to train and evaluate my model. I used the Pandas library.

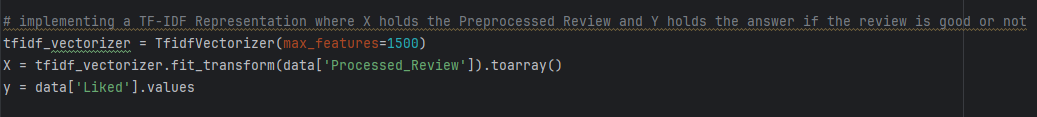
Text Preprocessing

A computer screen shot of a computer program

Description automatically generated

Next up is the preprocessing, it is a vital method used up to clean up the data and make it suitable for machine learning, I used a number of ways to preprocess the text, I removed non alphabetic symbols and characters from the data as well as converting the entire text into lowercase, I used Stemming to basically reduce the words into their base shape, I also used tokenization and lemmatization.

Bag-of-Words Representation



I implemented it using the TF-IDF representation and gave it a maximum number of features of 1500 to allow the model to capture the important words well.

Train-Test Split



I divided the dataset into training and testing.

Naïve Bayes Model and Predictions

A computer screen shot of a program

Description automatically generated

I then implemented a Naïve Bayes classifier and trained it on the preprocessed and TF-IDF training data. Following I made the model perform predictions on the test data and then printing the accuracy.

Confusion Matrix

A computer screen with colorful text

Description automatically generated

Lastly, I generated a confusion matrix to be able to clearly evaluate the performance of the model as a heatmap. The performance is evaluated in the form of true positives, true negatives, false positives, and false negatives.