Y19 Artificial Intelligence II
Deep Learning for Natural Language Processing

Oikonomou Stylianos - 1115201500243

Announced: November 8, 2022 $\|Due : November 29, 2022$

1 Preprocessing

In the preprocessing stage our goal is to remove as much noise as possible from the data sets that will interfere with the classification. The methods used to "clean up" the reviews are the following:

- Remove the left over HTML tags
- Remove words containing numbers
- Remove all non letter characters
- Clean up words containing apostrophe
- Remove multiple white spaces
- Turn all letters to lower case
- Remove stop words

The urls of the reviews were dropped from the data set as they could not provide significant context about whether a rating is positive or not.

The numbers 0 or 1 replaced the movie rating values to symbolise the negative and positive ratings respectively. Thus the classification turns into a binary problem based on those values.

2 Feature Selection

The features of the model were created by the CountVectorizer function which produced a sparse matrix containing the word counts for each word of the review.

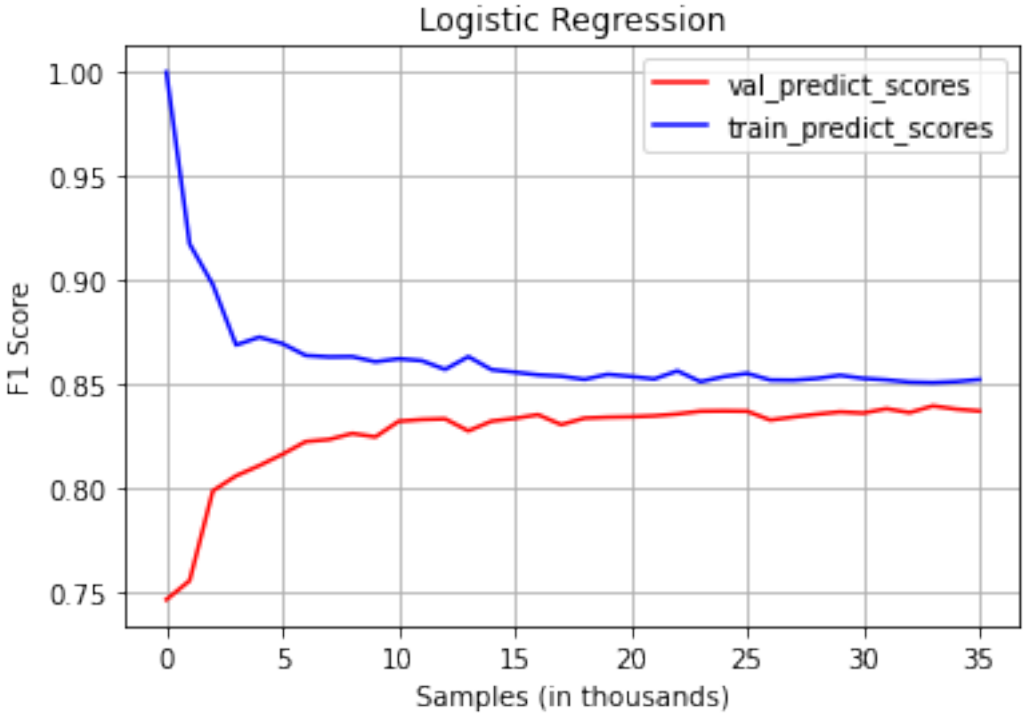
The max_features parameter of the function played an important role in the fine tuning of the model as it helped to solve overfitting by reducing the number of the features that were fed to the model.

The features were scaled by the StandardScaler function. Though unexpected the standardizing of the data did not change the final scores of the model.

Another scale function that was attempted was the MinMaxScaler but the function would not accept a sparse matrix as its input.

The data sets were split into two constant size subsets of 80% for the Train Set and 20% for the Validation Set. K-fold cross validation was also used to split the data but produced the same scores as the simple split method and was avoided due to the large time cost.

3 Results



The model reached an accuracy score of 84% and plateaued at around 17 thousand reviews. The gap between the learning curves of the Traning and Validation set scores was reduced to 0.015.