Insights and observations about the Data.

CodeLabsAcademy Study Case

WRITTEN BY:
DANIA ADIMI

09/03/2022

Summary of the notebook:

Insights, observations and performance

- The dataset **'imdb reviews'** contains labelled movie reviews (positive and negative).
- We can easily turn 'Tensorflow PrefetchDataset' into a dataframe in order to run some analysis.
- The resulting dataframe contains two columns:
  - **Review**, which contains the plain text of the review.
  - **Label**, 0 for a negative review and 1 for a positive review.
- We need to do text preprocessing before feeding it to the models and run some analysis, so we need to remove special chars, break lines, stop words, and do some stemming.
- After cleaning the data, we can extract the following insights:
  - The number of movies with positive reviews is which is slightly larger than the number of movies with negative reviews.
  - In average, positive reviews contain slightly more words then negative ones. However, negative reviews tend to be longer.
  - The words 'movie', 'film', 'one', and 'like' are most commun in both reviews. However, positive reviews have a high count of the following words: 'great', 'love', 'best', and negative reviews, on the other hand, have a high count of 'bad'.

Summary of the notebook:

Insights, observations and performance

- The task at hand is about sentiment analysis.
- In the predictive modeling section, we've different classification models.
- We observed that the LinearSVC model performed better than both Logistic regression and Gaussian classifier.
- We can also use neural network in order to classify our data. As demonstrated in the notebook.
- We still can improve the accuracy of the models by preprocessing the data again and explore other models, the process is always iterative.