Allen kombasseril
z5232188

Report

1. *Evaluation of your stacking model on the test data.*

I have obtained a F1 score of **0.7483312619309965**, on the test set .

2. *How would you improve the performance (e.g., F1) of the stacking model.*

Initially my performance was about 0.7483312619309965, from there I could get the performance on the testset upto 0.7968917361766614.

I did this by replacing tokenizer with regextokenizer and i also added stopwords, i will explain what these are below:

The way we tokenize our data is important, the normal tokenizer only splits the words when there is a whitespace, and therefore some words like “hello” and “hello!” is treated differently. To avoid this , I used regextokenizer to split words on commas, question mark, exclamation mark and full stop in addition to whitespaces.

Another method I used is by including stopwords in the pipeline, which removes common words like “the” , “is”, “are” , “and” etc from the features , which showed a slight improvement in performance.

Another method is to run grid_search_cv to find the best hyperparameters for our models.