Preprocessing:
Using some python magic to read the json file of tweets
Only the tweet.text column was used.
Some irrelevant yet abundant symbols such as “@” were removed.
Feature engineering
A simple bag of words approach was used because the amount of data is too large for processing in a reasonable amount of time
Model:
The machine learning algorithm used was the xgboost model
Result:
The result was pretty mediocre due to the limitation of the approach.
Insights:
I would have considered using a pre-trained word2vec model if I had time. (Difficult to use on kaggle kernels)
However, I personally think that using a pre-trained model is against the spirit of the competition. That’s why I didn’t consider this approach initially.
Trying to construct a word2vec model from the given data would have probably had worse performance due the nature of social media posts