Text Classification

you will do experiments on text classification. you're challenged to build a multi-headed model that's capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate text present in social media/document posts. You can download the dataset from the following link.

Instruction:

1. Build a classification model and compare it with different text based feature methods like(raw text,BOW, embeddings). List the features set which gave better results with Proper Intuition.
2. Evaluate the model on test set with average F1 score as metric
3. Provide top 20 features(phrases) for your class prediction i.e(Top Phrases which is responsible for your class prediction to particular class)

Output:

1. Result.csv

Sample

Text	Classes	Top_Phrases
Your absurd edits on great white shark was total vandalism and was very sexual	[toxic,obscene]	[absurd,sexual,total vandalism….]
i'm going to keep posting the stuff u deleted until this fucking site closes down have fun u stupid ass bitch don't ever delete anything fuckin hore like i said before go to hell	[toxic,obscene,insult]	[stupid ass bitch, fucking hore, go to hell ...]

2. Jupyter/Collab Notebook or Python code of training and testing module.

Note:

1. Compressed your submission and please put necessary comments in your code.
2. Please provide all the dependencies needed to test the code.
3. Attach the standard doc link in readme if any pretrained model is used.

Please make necessary assumptions if needed and mention that in your README