Term Frequency or (Bag of words model) is used initially: But since our documents/text has various sizes more weightage is given to longer documents (30-5295).

TF-Inverse Document frequency: (number of terms in documents/number of documents): usually log(1) worst case scenario indicating the most common words and using this we take only top features(should be medium so as to include rare words) to reduce model complexity, overfitting and decrease bias

Pre-processing like lowercase, lemmatization, stop words, n-gram (combination of words), stop words etc is used to increase accuracy without it the accuracy is like 75 percent on validation data

Since its classification we don’t care about semantics

Naïve bayes classifier is used since its simple model for classification and faster implementation. After testing on SVC, Logistic regression we found SVC is giving better results on validation data(96%) and so this is chosen as our classification model. From Confusion matrix we see that acq label is kind of overfitted. And data augmentation or better classification model like bert can be used.



