1. Regarding tokenization, I decided to filter out all punctuation and numbers, since they do not carry any important semantic information. Furthermore, I lowercased the words to making the data case-insensitive. A potential drawback of this kind of tokenization is that some words containing punctuation, e.g. “two-letter”, would not be considered as one word but rather as two separate terms. I also filtered out all words appearing less than 20 times because the classifier took a very long time to run when I tried running it with 15 and 10.

2. I used TruncatedSVD to perform dimensionality reduction on the data.

3. I chose GaussianNB as model 1 and DecisionTreeClassifier as model 2

4. Accordingly to what we were asked to do for the assignment, I evaluated the two classifiers with unreduced features, with 50% of the available features, but also 25%, 10% and 5%. The results can be seen in the following table:

dimensions/classifier	Model1(GaussianNB)	Model2(DecisionTreeClassifier)
Unreduced features (around 10000 dimensions)	accuracy score: 0.7273209549071618 , precision score: 0.7299866978124553 , recall score: 0.7273209549071618 , F1 score: 0.7278183279218778	accuracy score: 0.5954907161803713 , precision score: 0.6015799072436602 , recall score: 0.5954907161803713 , F1 score: 0.5975260205471508
50 % dimensionality reduction (5000)	accuracy score: 0.12625994694960213 , precision score: 0.24751241359831455 , recall score: 0.12625994694960213 , F1 score: 0.09999460244227441	accuracy score: 0.15517241379310345 , precision score: 0.15835235417540428 , recall score: 0.15517241379310345 , F1 score: 0.15611481246916845
25% dimensionality reduction (2500)	accuracy score: 0.13262599469496023 , precision score: 0.3318482818922374 , recall score: 0.13262599469496023 , F1 score: 0.11386867939558755	accuracy score: 0.17082228116710876 , precision score: 0.17292292842113355 , recall score: 0.17082228116710876 , F1 score: 0.17096263691143043
10 % dimensionality reduction (1000)	accuracy score: 0.15145888594164456 , precision score: 0.3589768312620909 , recall score: 0.15145888594164456 , F1 score: 0.1417246464135361	accuracy score: 0.1856763925729443 , precision score: 0.18711763463009812 , recall score: 0.1856763925729443 , F1 score: 0.18584036821108543
5% dimensionality reduction (500)	accuracy score: 0.1687002652519894 ,	accuracy score: 0.17453580901856763 ,

	precision score: 0.33173284674987186 , recall score: 0.1687002652519894 , F1 score: 0.15459802053950203	precision score: 0.17885269257026057 , recall score: 0.17453580901856763 , F1 score: 0.17603229311925103

Explanation

Without performing dimensionality reduction, the GaussianNB classifier performs significantly better than the DecisionTreeClassifier. However, after performing dimensionality reduction of the features, the scores have a huge decline. For the Gaussian classifier, we have a reduction of around 62% for accuracy, recall and F1 score. Only the precision score that the Gaussian classifier gives is a bit higher (0.26 with 50% dimensionality reduction, 0.28 with 25%, 0.35 with 10% and 0.32 with 5%). Dimensionality reduction of the features with the DecisionTreeClassifier also lowers the scores by around 45%. No matter how many dimensions the feature table has, its accuracy, its precision, its recall and its F1 scores are all between 0,14 and 0,18, and do not differ as greatly as they do with the GaussianNB classifier.

Part Bonus:

I used PCA to perform dimensionality reduction of the features. I also changed the least amount of times a word has to appear in the corpus for it to be a feature from 20 to 15. The scores with reduced and unreduced features do not differ almost at all when we compare it with the scores that we get with TruncatedSVD.

	Model1(GaussianNB)	Model2(DecisionTreeClassifier)
Unreduced features (around 12000 dimensions)	accuracy score: 0.7453580901856764 , precision score: 0.7464330779014219 , recall score: 0.7453580901856764 , F1 score: 0.7452110429913603	accuracy score: 0.596816976127321 , precision score: 0.6040121781448392 , recall score: 0.596816976127321 , F1 score: 0.5991956367425268
50 % dimensionality reduction (6000)	accuracy score: 0.12917771883289125 , precision score: 0.26378965654247044 , recall score: 0.12917771883289125 , F1 score: 0.10410958674929732	accuracy score: 0.14880636604774536 , precision score: 0.1496774220718922 , recall score: 0.14880636604774536 , F1 score: 0.1487541534222727

25% dimensionality reduction (3000)	accuracy score: 0.12811671087533155 , precision score: 0.2804637504147577 , recall score: 0.12811671087533155 , F1 score: 0.10440947426912392	accuracy score: 0.1636604774535809 , precision score: 0.169107740052937 , recall score: 0.1636604774535809 , F1 score: 0.1655061335223497
10 % dimensionality reduction (1200)	accuracy score: 0.1572944297082228 , precision score: 0.3587381547301795 , recall score: 0.1572944297082228 , F1 score: 0.14627873507115846	accuracy score: 0.1673740053050398 , precision score: 0.17122728386653488 , recall score: 0.1673740053050398 , F1 score: 0.168742992689235
5% dimensionality reduction (600)	accuracy score: 0.16790450928381964 , precision score: 0.3233551033654175 , recall score: 0.16790450928381964 , F1 score: 0.15405774233906228	accuracy score: 0.18143236074270558 , precision score: 0.18454889454752627 , recall score: 0.18143236074270558 , F1 score: 0.18239577690438516