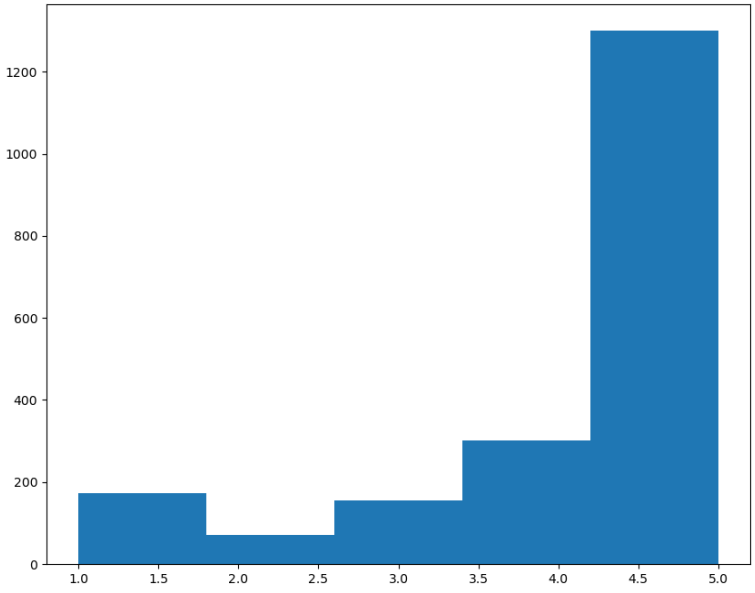
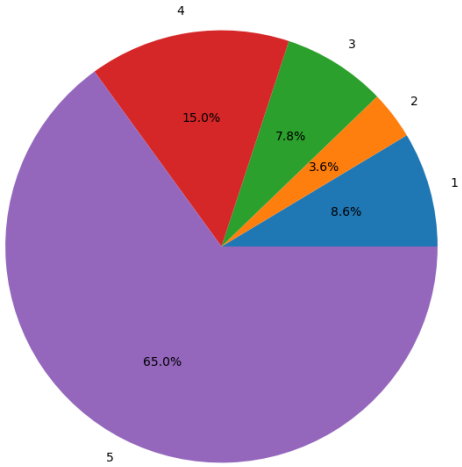
Assignment 2 Report

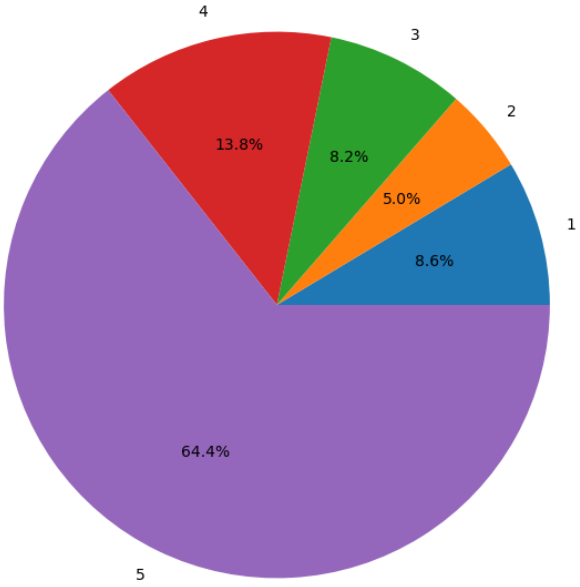
1. Data analysis

   First, let's analyze the rating are in the original train data, and what is the proportion.

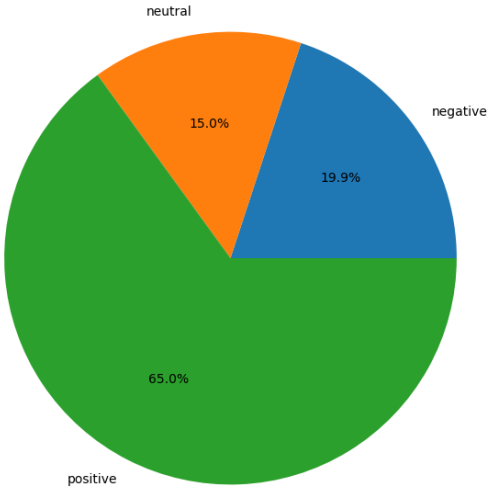




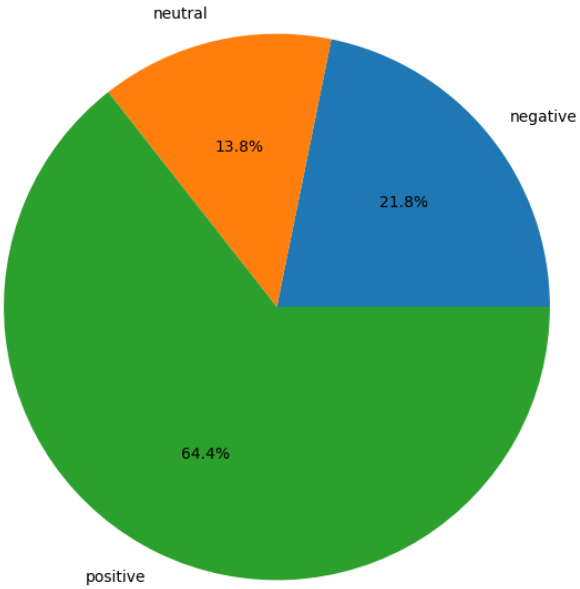
let's analyze the rating are in the original test data, and what is the proportion.



Then, we convert the rating into a sentiment and analyze it again. In train data.



In test data.



The frequently occurring words samples are:

your	999	would	990	won	982
you	998	worth	989	without	981
yet	997	world	988	with	980
yeast	996	works	987	wish	979
years	995	worked	986	wine	978
year	994	work	985	wife	976
wrong	992	word	984	why	975
wouldn	991	wonderful	983	whole	974

2. Develop Decision Tree models for training and testing: (a) with the 1% stopping criterion (the standard model), and (b) without the 1% stopping criterion.

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

In order to prevent overfitting, decision trees usually take pruning operations. If the number of samples on a branch and leaf is less than a certain threshold, the branch and leaf will be pruned. The following table shows the appropriate pruning. The performance of the model can be improved.

Evaluation	stopping	accuracy	Micro-precisio	Macro-precisio	Micro-recall	Macro-recall	Micro-F1	Macro-F1

	criterion		n	n				
scenario 1	1%	0.639	0.639	0.303	0.303	0.639	0.254	0.146
	not 1%	0.536	0.536	0.254	0.536	0.239	0.143	0.126
scenario 2	1%	0.658	0.658	0.479	0.658	0.419	0.648	0.315
	not 1%	0.535	0.535	0.367	0.535	0.357	0.535	0.598

3. Develop BNB and MNB models from the training set using: (a) the whole vocabulary (standard models), and (b) the most frequent 1000 words from the vocabulary, as defined using scikit-learn Count Vectorizer, after preprocessing by removing "junk" characters.

Extracting the first 1000 words of frequency can effectively prevent overfitting. At the same time, because the training set has only 2000 samples of data, taking the first 1000 words can prevent the phenomenon of dimension disaster.

From the final result, using the first 1000 words of frequency can indeed effectively improve the performance of the model, proving that the dimensional disaster does exist in the original model, and dimensionality reduction can effectively suppress the dimensional disaster.

model	Evaluation	vocabulary	accuracy	Micro-precision	Macro-precision	Micro-recall	Macro-recall	Micro-F1	Macro-F1
BNB	scenario 1	1000	0.639	0.639	0.303	0.303	0.639	0.254	0.146
		whole	0.536	0.536	0.254	0.536	0.239	0.143	0.126
	scenario 2	1000	0.658	0.658	0.479	0.658	0.419	0.648	0.315
		whole	0.535	0.535	0.367	0.535	0.357	0.535	0.598
BNB	scenario 1	1000	0.641	0.641	0.301	0.301	0.641	0.257	0.146
		whole	0.534	0.534	0.252	0.532	0.236	0.142	0.125
	scenario 2	1000	0.657	0.657	0.474	0.656	0.417	0.649	0.311
		whole	0.53	0.53	0.364	0.53	0.35	0.53	0.59

4. Evaluate the effect of preprocessing for the three standard models by comparing models developed with: (a) only the preprocessing described above (standard models), and (b) applying, in addition, Porter stemming using NLTK then English stop word removal using scikit-learn Count Vectorizer.

Using the NLTK function can effectively remove unnecessary words in a sentence, remove redundant and useless information, and make the model more expressive.

It can also be seen from the table below that after using NLTK, the effect of the model is indeed improved, which proves that some useless redundant information will indeed affect the accuracy of the model.

model	Evaluation	NLTK	accuracy	Micro-precision	Macro-precision	Micro-recall	Macro-recall	Micro-F1	Macro-F1
DT	scenario 1	not	0.624	0.624	0.302	0.624	0.254	0.148	0.132
		use	0.648	0.648	0.316	0.648	0.235	0.117	0.089
	scenario 2	not	0.635	0.635	0.478	0.635	0.416	0.635	0.315
		use	0.639	0.639	0.406	0.635	0.406	0.639	0.299

BNB	scenario 1	not	0.578	0.578	0.346	0.598	0.316	0.261	0.31
		use	0.586	0.586	0.333	0.586	0.176	0.179	0.145
	scenario 2	not	0.628	0.628	0.479	0.628	0.521	0.628	0.348
		use	0.607	0.607	0.445	0.608	0.418	0.609	0.304
MNB	scenario 1	not	0.603	0.603	0.358	0.609	0.337	0.264	0.235
		use	0.597	0.597	0.335	0.597	0.308	0.206	0.2
	scenario 2	not	0.663	0.663	0.548	0.663	0.534	0.663	0.401
		use	0.648	0.648	0.485	0.648	0.465	0.648	0.352

5. Evaluate the effect of converting all letters to lower case for the three standard models by comparing models with: (a) no conversion to lower case, and (b) all input text converted to lower case.

Changing all words to lowercase will reduce the types of words, and at the same time, the number of each word may increase, which enhances the aggregation ability of the model, so that the model does not need to fit more dimensional data.

According to the results of the model in the table below, it can be known that after changing all words to lowercase, the effect of the model has not been improved, but has declined. It can be seen that the case of words is also a very important feature of the sample.

model	Evaluation	lower case	accuracy	Micro-precision	Macro-precision	Micro-recall	Macro-recall	Micro-F1	Macro-F1
DT	scenario 1	no	0.628	0.628	0.295	0.628	0.257	0.146	0.2
		yes	0.626	0.626	0.258	0.626	0.23	0.102	0.08
	scenario 2	no	0.639	0.639	0.485	0.639	0.415	0.639	0.3
		yes	0.614	0.614	0.416	0.614	0.402	0.614	0.302
BNB	scenario 1	no	0.627	0.627	0.306	0.627	0.254	0.146	0.12
		yes	0.583	0.583	0.379	0.583	0.326	0.231	0.233
	scenario 2	no	0.637	0.637	0.458	0.637	0.416	0.637	0.31
		yes	0.626	0.626	0.482	0.626	0.471	0.626	0.352
0.351MNB	scenario 1	no	0.626	0.626	0.302	0.626	0.251	0.143	0.1
		yes	0.616	0.616	0.354	0.616	0.358	0.274	0.2
	scenario 2	no	0.636	0.636	0.485	0.636	0.415	0.636	0.312
		yes	0.681	0.681	0.561	0.681	0.553	0.681	0.43

6. Describe your chosen “best” method for rating prediction. Give new experimental results for your method trained on the training set of 2000 reviews and tested on the test set of 500 reviews. Explain how this experimental evaluation justifies your choice of model, including settings and parameters, against a range of alternatives. Provide new experiments and justifications: do not just refer to previous answers.

SVC has a strong ability to deal with classification problems. It can be seen from the table below that SVC is indeed the best performing model among the four models.

model	accuracy	Micro-precision	Macro-precision	Micro-recall	Macro-recall	Micro-F1	Macro-F1
DT	0.639	0.639	0.303	0.303	0.639	0.254	0.146
BNB	0.536	0.536	0.254	0.536	0.239	0.143	0.126
MNB	0.658	0.658	0.479	0.658	0.419	0.648	0.315
my	0.66	0.66	0.32	0.66	0.304	0.211	0.2