Assignment 2 Report

1. First, let's analyze the data, 2 stars is the minimum amount, 5 stars has the most number, more than the rest combined. It can be seen that most customers have a good evaluation of the product. It can also explain from the side that the quality of the product is very good.

Rating	1	2	3	4	5
Train Set	8.6%	3.6%	7.75%	15.05%	65%
Test Set	8.6%	5%	8.2%	13.8%	64.4%

Table1 Proportion of each class

Then, we convert 5 classes in scenario 1 to 3 classes in scenario 2, where the rating are combined into a “sentiment”, where 1, 2 or 3 is negative, 4 is neutral and 5 is positive. As can be seen from the table below, most of the users' attitudes towards product use are positive. Negative and neutral quantitative attitudes are not much different.

Sentiment	negative	neutral	positive
Train Set	19.95%	15.05%	65%
Test Set	21.8%	13.8%	64.4%

Table2 Proportion of each sentiment

The 15 most frequently occurring words in the training set. The frequencies are all less than 1000 and greater than 980.

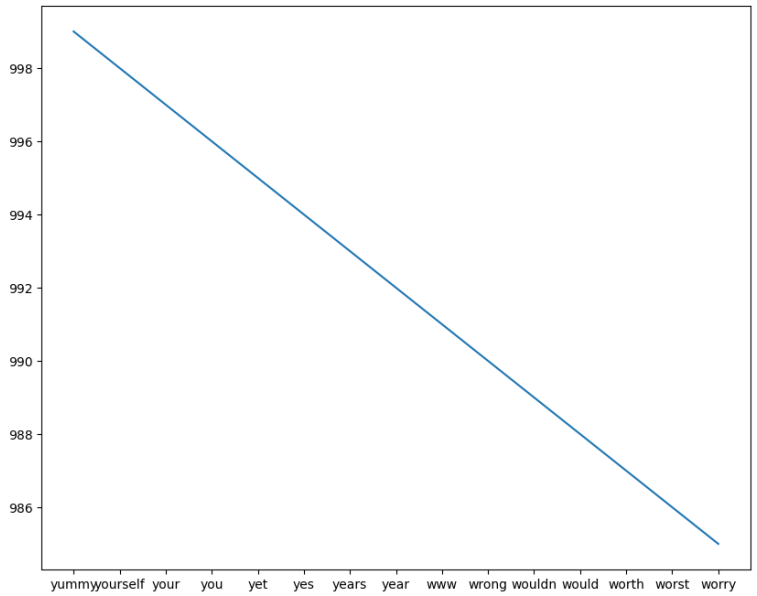

Fig1 The word with the highest word frequency in the training set.

The 15 most frequently occurring words in the test set. The frequencies are all less than 1000 and greater than 980.

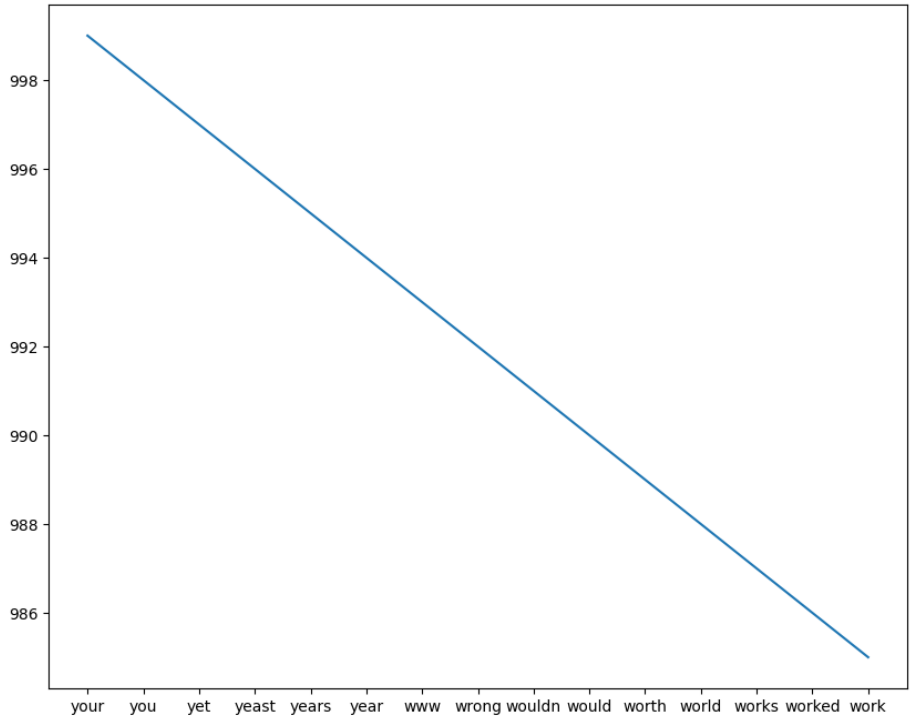

Fig2 The word with the highest word frequency in the test set.

2. Develop Decision Tree models for training and testing for scenario 1.

Stop criterion	With 1%	Without 1%
accuracy	0.636	0.532
Micro- precision	0.636	0.532
Macro- precision	0.298	0.241
Micro-recall	0.636	0.532
Macro-recall	0.253	0.238
Micro-F1	0.116	0.142
Macro-F1	0.142	0.119

Table3 Evaluation of DT in scenario 1

Develop Decision Tree models for training and testing for scenario 2.

Stop criterion	With 1%	Without 1%
accuracy	0.646	0.532
Micro- precision	0.646	0.532
Macro- precision	0.4728	0.359
Micro-recall	0.646	0.532
Macro-recall	0.416	0.358

Micro-F1	0.646	0.532
Macro-F1	0.312	0.268

Table4 Evaluation of DT in scenario 2

When we set 1% stop criterion, the min samples leaf parameter will be set as 20, it means set the number of samples on the leaves of each decision tree, less than the set value will be pruned. It can effectively prevent overfitting and improve the prediction accuracy of the test set. In scenarios 1 and 2, the accuracy, Micro- precision, Micro-recall score are the same. In scenarios 1, we divided the dataset into five parts, the difficulty of the prediction increases, and the predicted score is lower. In scenarios 2, we divided the dataset into three parts, the difficulty of prediction decreased and the predicted score increased.

3. BNB model prediction score in scenarios 1.

Training set using	the whole vocabulary	the most frequent 1000 words
accuracy	0.626	0.576
Micro- precision	0.626	0.576
Macro- precision	0.318	0.348
Micro-recall	0.626	0.576
Macro-recall	0.218	0.302
Micro-F1	0.076	0.194
Macro-F1	0.058	0.2

Table5 Evaluation of BNB in scenarios 1

BNB model prediction score in scenarios 2.

Training set using	the whole vocabulary	the most frequent 1000 words
accuracy	0.634	0.624
Micro- precision	0.634	0.624
Macro- precision	0.396	0.481
Micro-recall	0.634	0.624
Macro-recall	0.364	0.465
Micro-F1	0.634	0.624
Macro-F1	0.255	0.353

Table6 Evaluation of BNB in scenarios 2

It can be seen from the above table that the accuracy of the whole vocabulary is higher than that of the most frequent 1000 words, the reason may be that 1000 features cannot represent most of the features of the dataset well. As a result, the model cannot learn the correct classification ability and the prediction score is low. For accuracy, precision, recall, the score in scenarios 1 is higher than scenarios 1, but the F1 score is lower. It should be caused by the different number of categories.

MNB model prediction score in scenarios 1.

Training set using	the whole vocabulary	the most frequent 1000 words
accuracy	0.64	0.602
Micro- precision	0.64	0.602

Macro- precision	0.229	0.351
Micro-recall	0.64	0.602
Macro-recall	0.203	0.335
Micro-F1	0.011	0.258
Macro-F1	0.011	0.228

Table7 Evaluation of MNB in scenarios 1

MNB model prediction score in scenarios 2.

Training set using	the whole vocabulary	the most frequent 1000 words
accuracy	0.672	0.662
Micro- precision	0.672	0.662
Macro- precision	0.479	0.544
Micro-recall	0.672	0.662
Macro-recall	0.388	0.529
Micro-F1	0.672	0.662
Macro-F1	0.273	0.4

Table8 Evaluation of MNB in scenarios 2

It can be seen from the above table that the accuracy of the whole vocabulary is higher than that of the most frequent 1000 words, the reason may be that 1000 features cannot represent most of the features of the dataset well. As a result, the model cannot learn the correct classification ability and the prediction score is low. For accuracy, precision, recall, the score in scenarios 1 is higher than scenarios 1, but the F1 score is lower. It should be caused by the different number of categories.

4. DT model use or not use NLTK .

DT model prediction score in scenarios 1.

Training set	Not use NLTK	Use NLTK
accuracy	0.626	0.646
Micro- precision	0.626	0.646
Macro- precision	0.291	0.318
Micro-recall	0.626	0.646
Macro-recall	0.254	0.233
Micro-F1	0.145	0.114
Macro-F1	0.12	0.081

Table9 Evaluation of DT in scenarios 1

DT model prediction score in scenarios 2.

Training set	Not use NLTK	Use NLTK
accuracy	0.636	0.634
Micro- precision	0.636	0.634
Macro- precision	0.457	0.4
Micro-recall	0.636	0.634
Macro-recall	0.415	0.399
Micro-F1	0.636	0.634

Macro-F1	0.31	0.288

Table10 Evaluation of DT in scenarios 2

BNB model prediction score in scenarios 1.

Training set	Not use NLTK	Use NLTK
accuracy	0.576	0.582
Micro- precision	0.576	0.582
Macro- precision	0.345	0.321
Micro-recall	0.576	0.582
Macro-recall	0.304	0.278
Micro-F1	0.199	0.179
Macro-F1	0.199	0.175

Table11 Evaluation of BNB in scenarios 1

BNB model prediction score in scenarios 2.

Training set	Not use NLTK	Use NLTK
accuracy	0.62	0.604
Micro- precision	0.62	0.604
Macro- precision	0.473	0.441
Micro-recall	0.62	0.604
Macro-recall	0.459	0.413
Micro-F1	0.62	0.604
Macro-F1	0.348	0.314

Table12 Evaluation of BNB in scenarios 2

MNB model prediction score in scenarios 1.

Training set	Not use NLTK	Use NLTK
accuracy	0.602	0.594
Micro- precision	0.602	0.594
Macro- precision	0.354	0.336
Micro-recall	0.602	0.594
Macro-recall	0.339	0.295
Micro-F1	0.261	0.2
Macro-F1	0.232	0.192

Table13 Evaluation of MNB in scenarios 1

MNB model prediction score in scenarios 2.

Training set	Not use NLTK	Use NLTK
accuracy	0.662	0.644
Micro- precision	0.662	0.644
Macro- precision	0.542	0.476
Micro-recall	0.662	0.644
Macro-recall	0.531	0.463

Micro-F1	0.662	0.644
Macro-F1	0.4	0.35

Table14 Evaluation of MNB in scenarios 2

Through the above three tables, comparing the three models, we can see that the use of NLTK improves the prediction results of the model. It can help us weed out some words with irrelevant mood and emotion, prevent the model from being influenced by these words to improve the accuracy of the model.

5. DT model whether uses lowercase.

DT model prediction score in scenarios 1.

Training set	no conversion to lower case	text converted to lower case
accuracy	0.626	0.624
Micro- precision	0.626	0.624
Macro- precision	0.291	0.273
Micro-recall	0.626	0.624
Macro-recall	0.254	0.23
Micro-F1	0.145	0.102
Macro-F1	0.12	0.08

Table15 Evaluation of DT in scenarios 1

DT model prediction score in scenarios 2.

Training set	no conversion to lower case	text converted to lower case
accuracy	0.636	0.604
Micro- precision	0.636	0.604
Macro- precision	0.457	0.413
Micro-recall	0.636	0.604
Macro-recall	0.415	0.392
Micro-F1	0.636	0.604
Macro-F1	0.31	0.292

Table16 Evaluation of DT in scenarios 2

BNB model prediction score in scenarios 1.

Training set	no conversion to lower case	text converted to lower case
accuracy	0.626	0.582
Micro- precision	0.626	0.582
Macro- precision	0.291	0.378
Micro-recall	0.626	0.582
Macro-recall	0.254	0.325
Micro-F1	0.145	0.23
Macro-F1	0.12	0.233

Table17 Evaluation of BNB in scenarios 1

BNB model prediction score in scenarios 2.

Training set	no conversion to lower case	text converted to lower case
accuracy	0.636	0.626
Micro- precision	0.636	0.626
Macro- precision	0.457	0.486
Micro-recall	0.636	0.626
Macro-recall	0.415	0.47
Micro-F1	0.636	0.626
Macro-F1	0.31	0.357

Table18 Evaluation of BNB in scenarios 2

MNB model prediction score in scenarios 1.

Training set	no conversion to lower case	text converted to lower case
accuracy	0.626	0.614
Micro- precision	0.626	0.614
Macro- precision	0.291	0.371
Micro-recall	0.626	0.614
Macro-recall	0.254	0.354
Micro-F1	0.145	0.274
Macro-F1	0.12	0.25

Table19 Evaluation of MNB in scenarios 1

MNB model prediction score in scenarios 2.

Training set	no conversion to lower case	text converted to lower case
accuracy	0.636	0.68
Micro- precision	0.636	0.68
Macro- precision	0.457	0.559
Micro-recall	0.636	0.68
Macro-recall	0.415	0.555
Micro-F1	0.636	0.68
Macro-F1	0.31	0.417

Table20 Evaluation of MNB in scenarios 2

Only the MNB model will improve the accuracy of converting to lowercase prediction in scenario 2, and the accuracy of the model will decrease in other cases. It can be seen that after all characters are lowercase, the feature expression ability of the text decreases, which is not conducive to the prediction of the model.

6. My model

In order to get better model predictions, I replaced the model with Linear SVC, which proved to be better at handling classification problems.

Training set	scenarios 1	scenarios 2
accuracy	0.66	0.67
Micro- precision	0.66	0.67

Macro- precision	0.32	0.324
Micro-recall	0.66	0.67
Macro-recall	0.304	0.308
Micro-F1	0.211	0.215
Macro-F1	0.2	0.204

Table19 Evaluation of Linear SVC