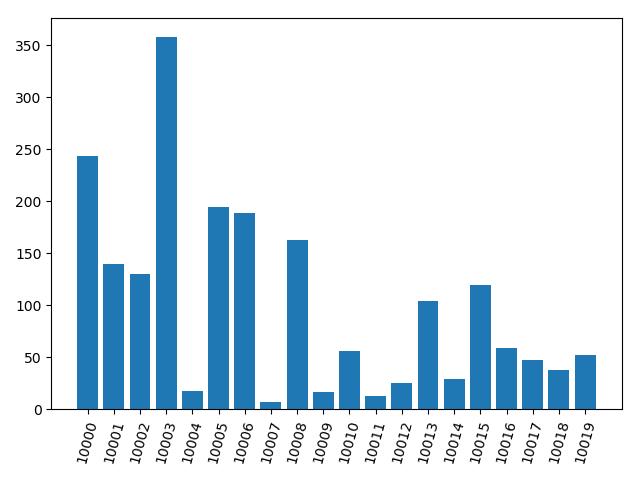
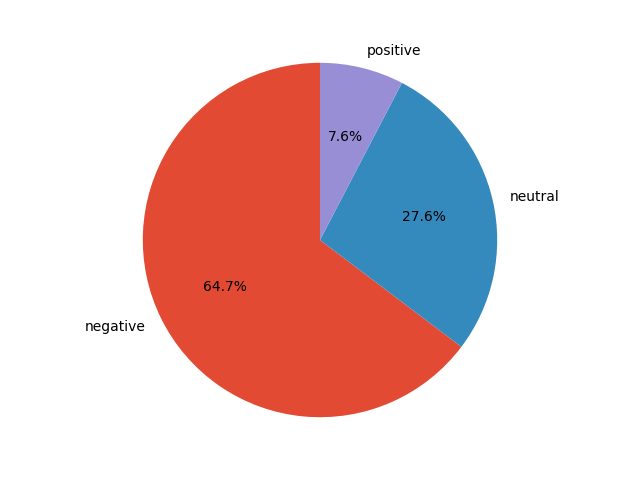
Assignment 2: Opinion Mining

1. **The frequency distributions for the sentiment and topic classes across the full dataset.**



**Figure 1.** The frequency distributions for the topic classes

From the figure we can see that these comments’ topics are not evenly distributed, but are concentrated on several topics, such as 10003 (Economic management), 10000 (corruption / governance), 10005 (healthcare / medicare) and 10006 (Social issues / marriage equality / religion). Those topics are issues that are widely concerned in daily life.



**Figure 2.** The frequency distribution for the sentiment classes

From the figure we can see that most comments are negative, only a few are positive.

1. **The baseline models’ performance and runtime vary with feature dimensions.**

Top N {100, 200, 300} words are selected as vocabulary, and using the tf-idf matrix as features. The metrics and runtime of corresponding model are shown in below (train dataset metrics / test dataset metrics):

Sentiment analysis:

300 words

200 words

100 words

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | Recall | Precision macro | Precision micro | F1\_score | Runtime |
| DT | 0.9713/0.65 | **0.9504/0.511** | 0.9769/0.5142 | 0.9713/0.65 | 0.9630/0.5108 | 0.030s |
| BNB | 0.7327/0.684 | 0.5974/0.506 | 0.6724/0.5402 | 0.7327/0.684 | **0.6247/0.5147** | 0.027s |
| MNB | **0.7013/0.696** | 0.4390/0.4061 | **0.8038/0.7477** | **0.7013/0.696** | 0.4494/ 0.4060 | **0.012s** |
|  | | | | | | |
| DT | 0.9913/0.636 | 0.9812/0.4851 | 0.9937/ 0.5002 | 0.9913/0.636 | 0.9873/0.4901 | 0.053s |
| BNB | 0.7527/0.72 | **0.6373/0.5402** | 0.6994/0.5987 | 0.7527/0.72 | **0.6615/0.5575** | 0.032s |
| MNB | **0.732/0.726** | 0.4923/0.4534 | **0.8245/0.7816** | **0.732/0.726** | 0.5246/0.4683 | **0.021s** |
|  | | | | | | |
| DT | 0.994/0.692 | **0.9843/0.5640** | 0.9951/ 0.5861 | 0.994/0.692 | 0.9896/0.5716 | 0.066s |
| BNB | **0.7767/0.736** | 0.6645/0.5576 | 0.7311/0.666 | **0.7767/0.736** | **0.6902/0.5775** | 0.078s |
| MNB | 0.7467/0.734 | 0.5151/0.4664 | **0.8402/0.7959** | 0.7467/0.734 | 0.5556/0.4891 | **0.024s** |

**Table 1.** Sentiment analysis: the metrics and runtime of corresponding

model with different vocabulary size.

From the Table 1, we can see that the Decision Tree (DT) model heavily overfitting on the dataset. When the vocabulary size used in feature processing is 300, the accuracy of 0.994 can be achieved on the training set, but on the test set, the accuracy is less than 0.7. This may be caused by the size of the dataset.

-

Compared with DT, there is no obvious overfitting in the two models of Bernoulli Naive Bayes (BNB) and Multinomial Naive Bayes (MNB).

When the feature dimension is small (vocabulary size is 100), most metrics of the MNB model are higher than the BNB. But when the feature dimension rises (vocabulary size 200 and 300), the performance of the BNB gradually surpasses the MNB.

From the perspective of runtime, runtime grows as the number of features increases. And the MNB runs much faster than the other two models (approximately a half of runtime).

Topic classification:

300 words

100 words

200 words

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | Recall | Precision macro | Precision micro | F1\_score | Runtime |
| DT | 0.8153/0.3384 | 0.7438/0.2153 | 0.8687/0.2167 | 0.8153/0.3384 | 0.7834/0.2109 | 0.078s |
| BNB | **0.4727/0.4146** | **0.2984/0.231** | 0.4347/0.2244 | **0.4727/0.4146** | **0.3097/0.2233** | 0.073s |
| MNB | 0.4247/0.3963 | 0.2125/0.1891 | **0.284/0.2397** | 0.4247/0.3963 | 0.2203/0.1901 | **0.026s** |
|  | | | | | | |
| DT | 0.8587/0.3232 | 0.8027/0.2305 | 0.9084/0.231 | 0.8587/0.3232 | 0.8390/0.2238 | 0.121s |
| BNB | **0.5707/0.4909** | **0.3709/0.2845** | **0.6253/0.2995** | **0.5707/0.4909** | **0.3891/0.2808** | 0.094s |
| MNB | 0.5013/0.4146 | 0.2562/0.2029 | 0.4415/0.2789 | 0.5013/0.4146 | 0.2651/0.2091 | **0.040s** |
|  | | | | | | |
| DT | 0.864/0.3689 | **0.822/0.331** | **0.9145/0.341** | **0.864/0.3689** | 0.846/0.3165 | 0.182s |
| BNB | **0.614/0.4665** | 0.3949/0.2571 | 0.591/0.272 | 0.4171/0.2538 | **0.6902/0.5775** | 0.073s |
| MNB | 0.5427/0.4177 | 0.2874/0.2055 | 0.4048/0.2952 | 0.3023/0.21 | 0.5556/0.4891 | **0.069s** |

**Table 2.** Topic classification: the metrics and runtime of corresponding

model with different vocabulary size.

From the Table 2, we can see the experiment results that similar to the sentiment analysis: overfitting of DT model, more stable performance of BNB and MNB models, and MNB is still the fastest model.

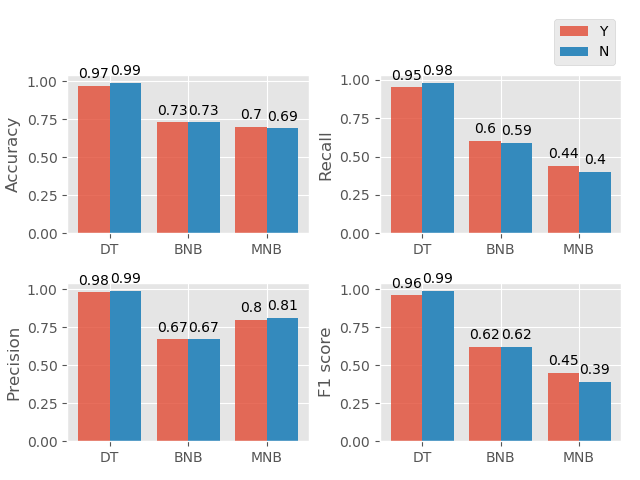
However, the difference is that the all baseline models have much lower metrics on the topic classification task. This should be because of there are 20 topics for all Twitter, and the distribution of topics is imbalance. But the performance of BNB is better than other two models.

**Conclusion:** The DT model is prone to overfitting in the case of small dataset, while NB models avoids overfitting due to the use of prior probabilities.

1. **Evaluate the standard models with respect to baseline predictors.**
2. **Evaluate the effect that preprocessing the input features.**

Sentiment analysis:

The vocabulary size is selected as 100.



**Figure 3.** The metrics on the sentiment analysis training dataset of three baseline models. Where Y presents preprocessing the input features with NLTK, N presents without preprocessing.

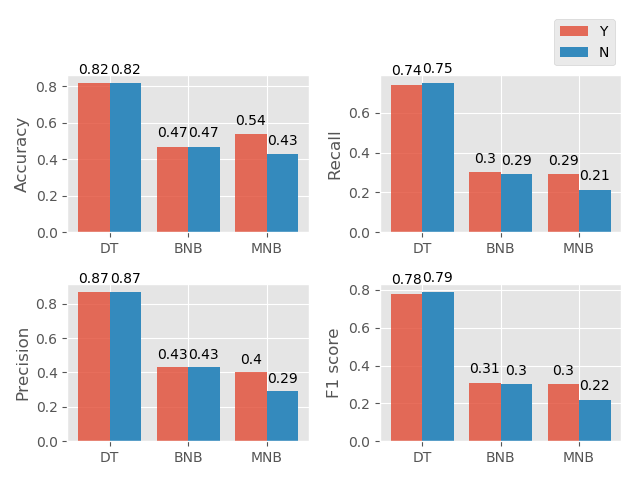


**Figure 4.** The metrics on the test dataset of three baseline models.

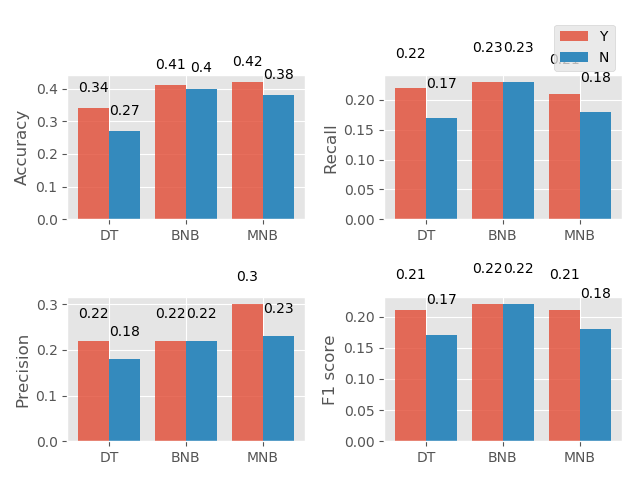
From the Figure 3, we can see that the metrics on the training dataset of baseline models are close with and without preprocessing. However, in the Figure 4, without the NLTK preprocessing, the metrics on the test dataset of DT and MNB drop a little. But the performance of BNB is better.

Topic classification:

The vocabulary size is selected as 100.



**Figure 5.** The metrics on the topic classification training dataset of three baseline models. Where Y presents preprocessing the input features with NLTK, N presents without preprocessing.



**Figure 6.** The metrics on the test dataset of three baseline models.

1. **Remove the neutral tweets and evaluate again.**