Limuelle Alamil

CMSC 197 ML Problem Set 4

Guide Questions:

1. What is the effect of removing stop words in terms of precision, recall, and accuracy? Show a plot or a table of these results.

   Answer: After removing the stop words in the email bodies, the accuracy and recall significantly increased while the precision stays almost the same. This implies that removing the stop words in the email bodies can increase the classifier's likelihood of correctly classifying an email as spam or ham (accuracy), as well as the likelihood that actual spam is predicted as spam (recall). However, the likelihood that a predicted spam is actually spam (precision) is not affected.
   Figure 1 compares the classification report of the NB classifier without and with stop words in the email body.

Dataset	Accuracy	Recall	Precision
no stop words	0.943973	0.942905	0.972206
with stop words	0.856002	0.802789	0.977102

Figure 1. Performance of classifiers when stop words are removed and retained.

2. Experiment with the number of words used for training.

   Answer: The number of words used for training affects the performance of the classifier. For instance, when the words used for training is limited only to words that have a frequency of more than 150, the accuracy, recall, and precision decreases slightly (see Figure 2). The report below shows that the scores have decreased compared to when the number of unique words is 10000 (see Figure 1).

```
Accuracy: 93.50%
Recall: 93.94%
Precision: 96.21%
```

   Figure 2. Performance of the classifier when the words are reduced

   A probable cause for this is that reducing the cardinality of the vocabulary would result in missing some spam words that are not frequent in the training set, thus, the reduced performance. Reducing the cardinality of the vocabulary, however, can decrease the training duration.

3. Discuss the results of the different parameters used for Lambda smoothing. Test it on 5 varying values of the λ (e.g. λ = 2.0, 1.0, 0.5, 0.1, 0.005). Evaluate performance metrics for each.

   This is the classification report when the classifier used a lambda value of 2.0.

   ```
   Accuracy: 94.33%
   Recall: 94.29%
   Precision: 97.12%
   ```

   This is the classification report when the classifier used a lambda value of 1.0.

   ```
   Accuracy: 94.40%
   Recall: 94.29%
   Precision: 97.22%
   ```

   This is the classification report when the classifier used a lambda value of 0.5.

   ```
   Accuracy: 94.42%
   Recall: 94.23%
   Precision: 97.32%
   ```

   This is the classification report when the classifier used a lambda value of 0.1.

   ```
   Accuracy: 94.39%
   Recall: 94.08%
   Precision: 97.41%
   ```

   This is the classification report when the classifier used a lambda value of 0.005.

   ```
   Accuracy: 94.24%
   Recall: 93.77%
   Precision: 97.49%
   ```

   This is the summary of the performance.

Dataset	Accuracy	Recall	Precision
0.005	94.24%	93.77%	97.49%
0.100	94.39%	94.08%	97.41%
0.500	94.42%	94.23%	97.32%
1.000	94.40%	94.29%	97.22%
2.000	94.33%	94.29%	97.12%

   Figure 3. Summary of the performance of classifiers with different lambda.

   There is no significant difference in the performance of the classifiers with varying lambda values. However, some trends are observable. Figure 4 shows that as the lambda increases, the recall increases while the precision decreases. Moreover,

accuracy is the highest when the lambda is 0.5. However, it starts to decrease as the lambda value increases or decreases from 0.5.

4. What are your recommendations to further improve the model?

   Answer: Here are a few recommendations to further improve the model:
   - A more standardized dataset can be used. The emails in the dataset used have varying encoding systems which make decoding more difficult. Due to this, I ended up using a more generalized way to decode which might have missed some words in some email bodies.
   - The email bodies can be cleaned further using techniques like stemming and lemmatization. Stemming can be used to group words based on their root stem and be referred to as one single word (e.g., standing, stands, and stood can be reduced to just stand). Lemmatization can be used to group words based on their root definition and be referred to as one sing word as well (e.g., stands and stand can be reduced to just stand). This can significantly reduce the cardinality of the vocabulary which can reduce training duration.