Lab 2: Sentiment Classification with Support Vector Machine

Group 9: Chen Yuan, Yuxuan Zhang, Veerle Uhl

1. Systematic Diagram of Sentiment Analysis Process

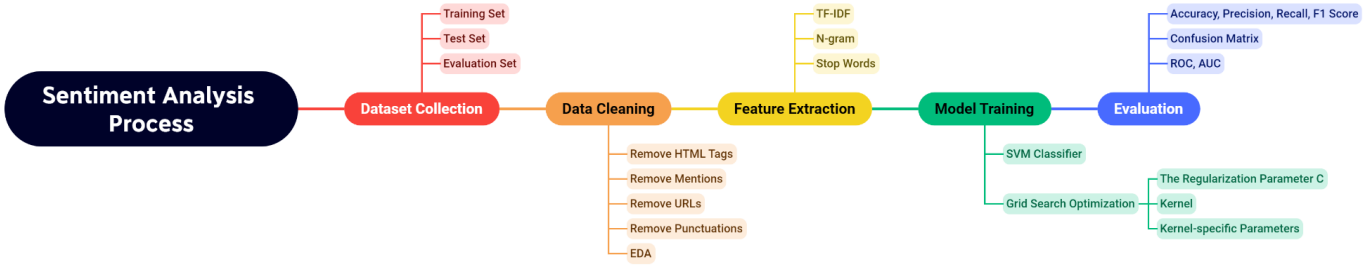

Fig 1. Systematic Diagram of Sentiment Classification with SVM

2. Data Cleaning & Feature Extraction

The data cleaning process is the same as the one in lab 1. HTML tags, mentions, URLs, and punctuation marks were considered irrelevant to sentiment and were therefore removed. The removed content was replaced with spaces to facilitate tokenization by the subsequent TF-IDF vectorizer. The TF-IDF vectorizer was applied again for feature extraction as in lab 1. With the experience from lab 1, this time special attention was paid to assessing the impact of the max features, the use of stop words, and the n-gram tokenization on the subsequent model's performance when constructing the TF-IDF vectorizer model. Based on the performance of subsequent models, we opted for a TF-IDF vectorizer with max features set to 5000, no use of stop words, and the consideration of only 1-gram as the optimal feature extraction approach. Furthermore, 1600 best features were selected in the end based on the Chi-squared test for the training, testing, and evaluation process.

3. Evaluation

The SVC was selected as the classifier, and a 5-fold cross-validation grid search optimization was applied to obtain the best hyperparameters. We compared the performance of the SVC with the linear kernel, the polynomial kernel, the RBF kernel, and the sigmoid kernel. In the end, we chose to use the RBF kernel with the gamma equal to 0.1 and the regularization parameter of the SVC equal to 15, as it gave the best accuracy of 0.8588 for the test set. Meanwhile, the accuracy for the evaluation set was 0.8524.

	precision	recall	f1-score	support
0	0.85	0.87	0.86	1252
1	0.86	0.85	0.86	1248
accuracy			0.86	2500
macro avg	0.86	0.86	0.86	2500
weighted avg	0.86	0.86	0.86	2500

	precision	recall	f1-score	support
0	0.85	0.86	0.85	2482
1	0.86	0.84	0.85	2518
accuracy			0.85	5000
macro avg	0.85	0.85	0.85	5000
weighted avg	0.85	0.85	0.85	5000

Fig 2. Classification Report of the SVM Method for the Test Set (Left) and the Evaluation Set (Right)

	precision	recall	f1-score	support
0	0.82	0.90	0.86	1252
1	0.89	0.80	0.84	1248
accuracy			0.85	2500
macro avg	0.85	0.85	0.85	2500
weighted avg	0.85	0.85	0.85	2500

	precision	recall	f1-score	support
0	0.78	0.94	0.85	2482
1	0.92	0.74	0.82	2518
accuracy			0.84	5000
macro avg	0.85	0.84	0.84	5000
weighted avg	0.85	0.84	0.84	5000

Fig 3. Classification Report of the Naive Bayesian Method for the Test Set (Left) and the Evaluation Set (Right)

4. Results

Overall, we achieved a good accuracy. To gain a clear understanding of the results, confusion matrices were plotted to visualize the numbers of true positives, true negatives, false positives, and false negatives. To compare the results from our Naive Bayesian classifier, the ROC curves were plotted and the AUC analysis was conducted. It can be seen that although the SVM classifier performs slightly better than the Naive Bayesian classifier in accuracy, the SVM classifier has a slightly worse AUC value, which means that in terms of the actual classification performance on the current test set and evaluation set, the SVM classifier is slightly inferior.

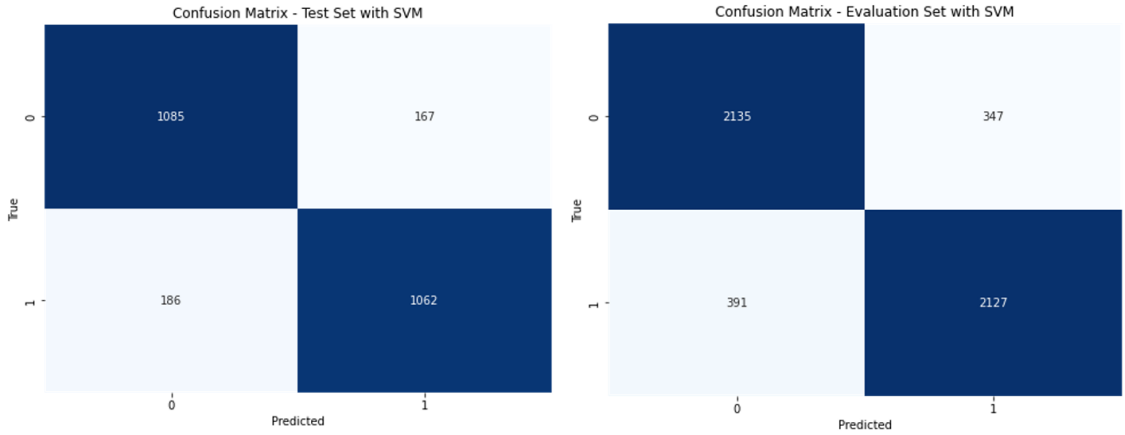

Fig 4. Confusion Matrices of the SVM Method for the Test Set (Left) and the Evaluation Set (Right)

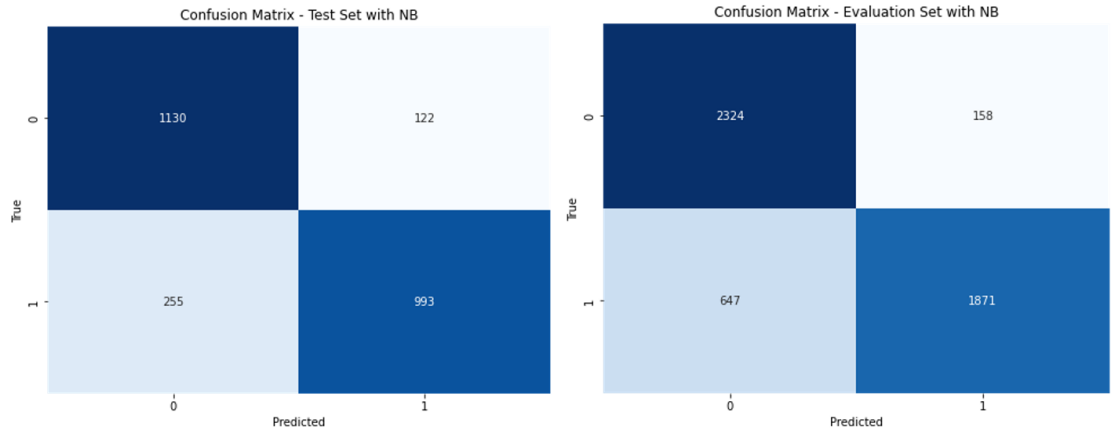

Fig 5. Confusion Matrices of the Naive Bayesian Method for the Test Set (Left) and the Evaluation Set (Right)

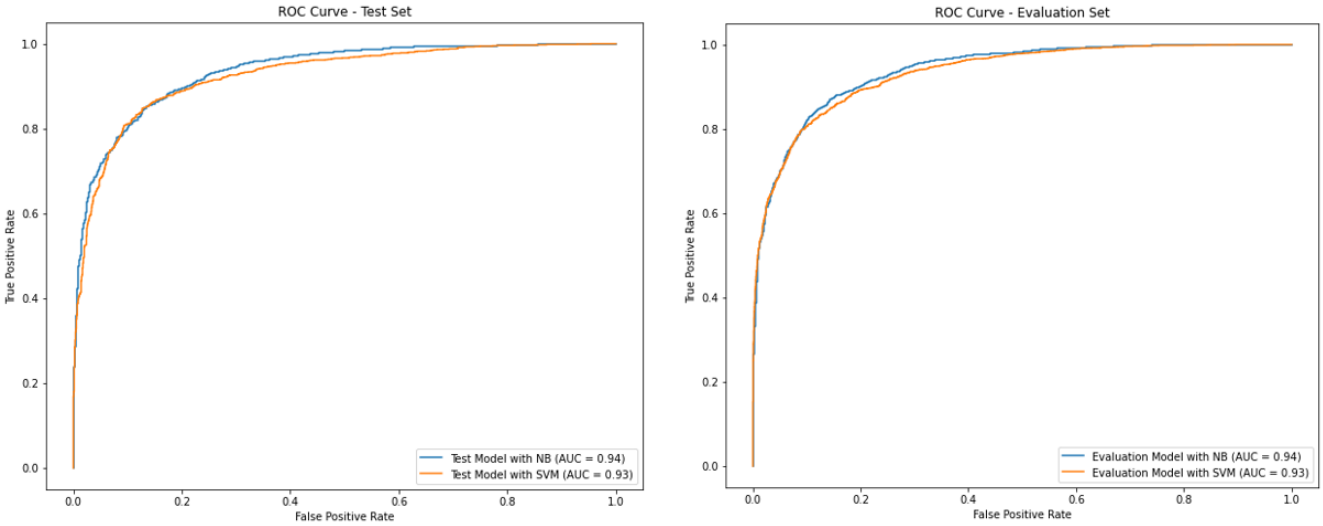

Fig 6. ROC & AUC of Both Methods for the Test Set (Left) and the Evaluation Set (Right)

5. Code

The code for our lab report can be found on github, using the following link:
https://github.com/KarlFran66/DM2583/tree/main/Lab%202