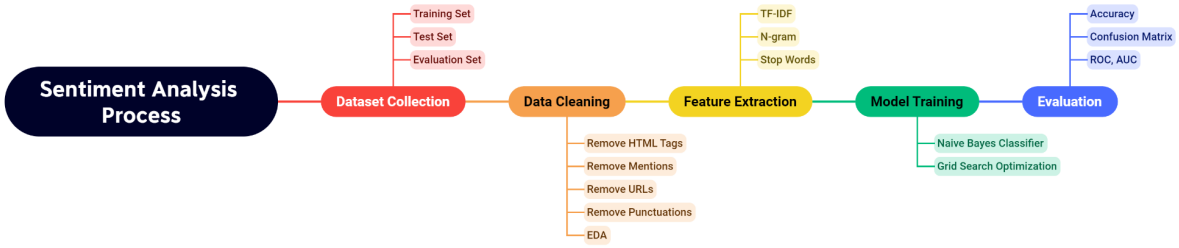
Lab 1: Sentiment Classification with Naïve Bayesian Classifier

Group 9: Chen Yuan, Yuxuan Zhang, Veerle Uhl

1. Systematic Diagram of Sentiment Analysis Process



2. Data Cleaning & Feature Extraction

During data cleaning, 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 for feature extraction as it converts text documents into a numerical feature matrix that can be used for machine learning models. The English stop words list was also used, and both 1-gram and 2-gram were considered.

3. Evaluation

The MultinomialNB was selected as the classifier, and a 5-fold cross-validation grid search optimization was applied to obtain the best alpha for the classifier as 2.9. We obtained the best performance (accuracy) for the test set as 0.8432 with these hyperparameters. Meanwhile, the accuracy for the evaluation set was 0.8388. However, we found an interesting phenomenon (which is called “with trial” in the diagrams) when we didn’t consider stop words, the accuracy for the test set slightly improved to 0.8492, with the accuracy for the evaluation set as 0.839.

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. Additionally, based on the ROC curve and AUC analysis, it can be seen that the classifier trained without considering stop words performed slightly better, but the difference was marginal. Therefore, we concluded that when using the Naive Bayes classifier to train on this dataset, the use of a standard English stop words list has minimal impact.