MIE 1624 - Tweet Sentiment Analysis

Yihao Du

1 Two classifiers comparison

In this homework, we will compare performance of two models - Naive Bayes and Logistic Regression on classification task. The classifier with better performance will be used in our further sentiment analysis. One technique of 10-fold cross-validation will be applied to suport our decision of classifiers. Two significant measurements, such as ROC curve and accuracy, will be used in our analysis on classifier performance.

1.1 Algorithm description

1.1.1 Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes Theorem. An important assumption in this algorithm is conditional independence in the fields of feature. The joint predict probability is formulated as

$$\begin{aligned} P(C_k|x_1, x_2, \dots, x_m) &\propto P(C_k, x_1, x_2, \dots, x_m) \\ &\propto P(C_k)P(x_1|C_k)P(x_2|C_k)\dots P(x_m|C_k) \\ &\propto P(C_k)\prod_{i=1}^{m} P(x_i|C_k) \end{aligned}$$

With this formula, the corresponding probabilistic classifier is able to be constructed respect the assignment of label C_k

$$y = \underset{k=1,2\dots,n}{\operatorname{argmax}}\, P(C_k)\prod_{i=1}^{m} P(x_i|C_k)$$

1.1.2 Logistic Regression

Logistic regression is an important regression model respect to the probability of each classification in the fields of label. With the assumption about independence in the fields of feature, the binary logistic regression is also be formulated as

$$\begin{aligned} P(Y=1|x_1, x_2, \dots, x_m) &= \frac{\exp(\beta_0 + \sum_{j=1}^{m}\beta_j x_j)}{1+\exp(\beta_0 + \sum_{j=1}^{m}\beta_j x_j)} \\ P(Y=0|x_1, x_2, \dots, x_m) &= \frac{1}{1+\exp(\beta_0 + \sum_{j=1}^{m}\beta_j x_j)} \end{aligned}$$

With this regression model, the probabilistic classifier is defined as

$$y = \underset{Y=0,1}{\operatorname{argmax}}\, P(Y|x_1, x_2, \dots, x_m)$$

1.2 Implementation details

The feature field for the classifiers needs to be determined before our model training. Here, we will use the frequency of words in the tweet as the feature in the classifier model, which is able to be obtained in terms of application bag-of-words methods in the training dataset.

In the next step, re-sampling technique is applied in the whole dataset to improve the estimation of the classification error, which greatly support the algorithm decision. Also, 10-fold cross-validation method is implemented to compare the two classifiers. In other words, nine tenths of the re-sampling dataset is used to train our classifier model, and the rest is used to test the model in last step.

1.3 The results of two classifies performance comparison

In order to select the suitable algorithm for our classification task, the results will be investigated respect to two measurements - *ROC curve* and *accuracy* based on 10-fold testing data.

ROC curve (receiver operating characteristic curve) is a great tool to evaluate the a binary classifier's performance. This curve is generated based on two key components, TPR (true positive rate) and FPR (false positive rate), with the following formula.

$$TPR = \frac{TP}{P} \qquad\qquad FPR = \frac{FP}{N}$$

Here, TPR indicates the proportion of the real positive elements that is labeled as positive by the classifier, while FPR indicates the proportion of the real negative elements that is labeled as positive by the classifier. An ideal classifier intends to maximize correct label (TPR) and minimize incorrect label (FPR). It is noticed that this goal can be achieved in a curve with a larger AUC (area under curve).

In order to visualize performance difference between two classifiers in terms of ROC curve, TPR and FPR have been averaged among all of 10 folds test data. As is shown in Figure 1, the red curve about Logistic Regression is associated with a larger AUC. It means this classifier performs better than Naive Bayes.

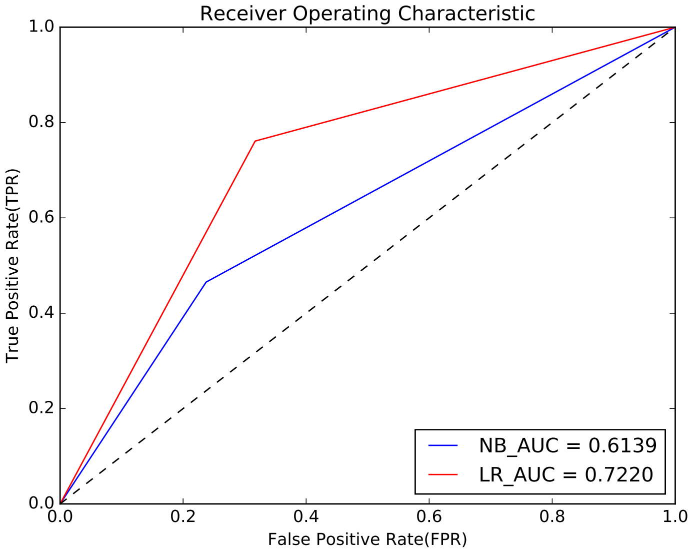

Figure 1: ROC curve for different classifiers

Another measure, *accuracy*, is used to describe total proportion of correctly labeled elements. An ideal model are supposed to lead to high accuracy in classification task. The follow boxplot is created based on accuracy in all of 10 folds test data. We can see that Logistic Regression is accompanied with higher mean and less variance in Figure 2. This plot implies the same conclusion with ROC curve that Logistic Regression has better performance than Naive Bayes. Therefore, *Logistic Regression* will be selected in our further classification task.

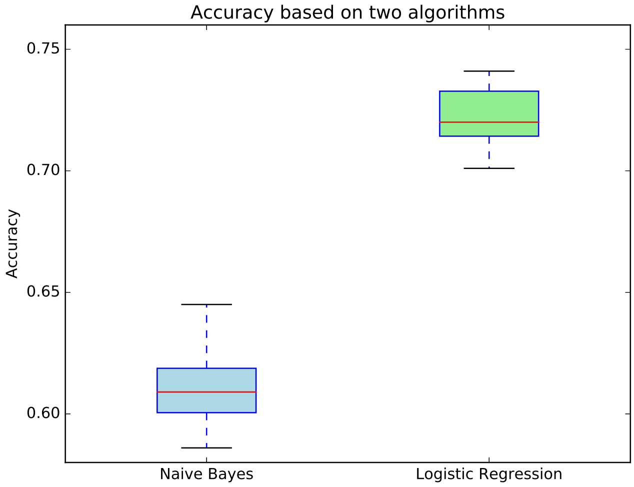

Figure 2: Accuracy of different classifiers in 10-fold cross-validation

2 Political sentiment analysis on the tweets

2.1 Total sentiment analysis

The following figure shows total sentiment of the tweets. It is very easy to see the positive tweet dominate in discussion about the 2015 federal election in Canada. We may conclude that people are content in the result of this election.

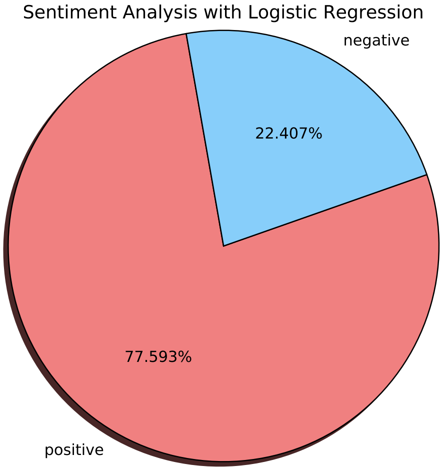

Figure 3: Sentiment analysis on the tweets

2.2 Political Sentiment analysis

Figure 4 shows the tweet topic composition respect to different political parties. We can notice that Liberal party was the most preferred party, besides other parties, mentioned in this discussion. It is interesting that rank of political party preference based on Figure 4 agrees on the result of the 2015 federal election.

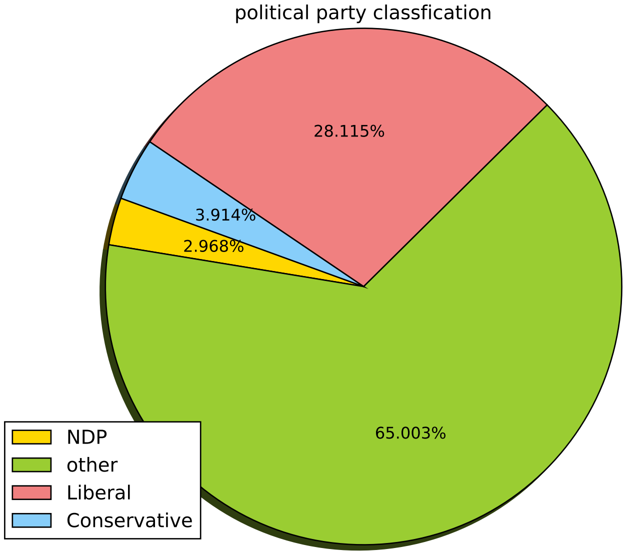

Figure 4: Political party composition

Figure 5 shows political sentiment in this federal election. We are able to observe that positive tweet also dominates dominates in discussion about all of political parties. This politically positive attitude provides further justification for the conclusion about sentiment in Section 2.1.

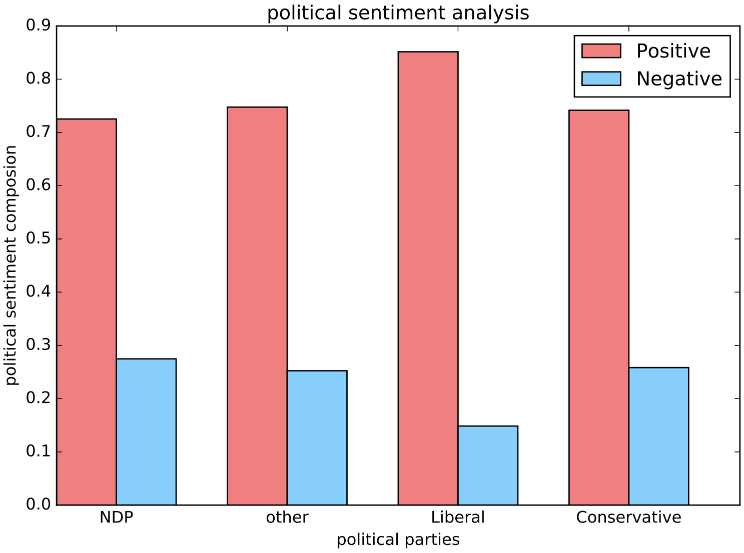

Figure 5: Political sentiment in different parties