COMP30018 Knowledge Technologies 2016 SM1 Project2 Report

Tessa(Hyeri) Song _ 597952

1. Introduction

1.1. The major problem and data sets involved

The major goal of this report is to solve a knowledge problem which requires data mining and machine learning. The main resource data to be manipulated is 50000 film reviews which is introduced by Maas et al (2011). Basically, the knowledge problem is to decide if films are 'good' or not based on the reviews. To decide this, it is essential to classify each review as either positive or negative toward its film. Among numerous classifiers, Naive Bayes and Random Forest has been chosen to accomplish this task and the evaluations and observations on the output will be described in detail in this report.

1.2. Overview of the classifiers

In order to apply the two classifiers mentioned above, each review will represent an instance and a bag of specific words will be chosen as 'attributes' and the frequency of the words will be values. The 'attributes' will play a crucial role to build models out of the classifiers to decide 'class' of each review which is either 'positive' or 'negative'.
The whole data sets are divided into two groups which are training and development sets and the former one is to be used to build models and the latter one to test and evaluate the models.

2. Building and Evaluating Models

2.1 Attributes

Selecting 'attributes' are a crucial part in building useful models. In this experiment, three different lists of attributes are used. The first list of attributes consists of 10 words which are the most frequent tokens in the collection. The second list is comprised of 200 words associated most with positive reviews according to the method introduced by Christopher Potts (2011) and additionally, other extra general 70 'good' words are also included to improve models. The last list is a combination between the words in the second list and 50 words associated most with negative reviews according to the same method above and extra 20 more general 'bad' words. For reference, general good and bad words are 'recommend', 'nice', 'hilarious', 'pleasant', 'surprising', 'funny', 'disappointing', 'stupid', 'boring', and so on. When choosing these words, ones with higher frequency in the collection were given priority.

2.2 Evaluation of models built by Naive Bayes

In order to evaluate each model for each list of attributes, precision, recall and confusion matrix are calculated. 'P' represents 'Positive' and 'N' represents 'Negative'.

1) The First list of attributes
   Time taken to build model : 0.05 seconds
   Accuracy : 52.192 %

		Predict	
		P	N
Actual	P	2486	10014
	N	1938	10562

Class	Precision	Recall
P	0.562	0.199
N	0.513	0.845

2) The Second list of attributes
Time taken to build model : 0.86 seconds
Accuracy : 64.556 %

		Predict	
		P	N
Actual	P	6751	5749
	N	3112	9388

Class	Precision	Recall
P	0.684	0.54
N	0.62	0.751

3) The third list of attributes
Time taken to build model : 1.15 seconds
Accuracy : 70.228 %

		Predict	
		P	N
Actual	P	9084	3416
	N	4027	8473

Class	Precision	Recall
P	0.693	0.727
N	0.713	0.678

2.3 **Evaluation of models built by Random Forest**

1) The First list of attributes
Time taken to build model : 18.76 seconds
Accuracy : 58.152 %

		Predict	
		P	N
Actual	P	7393	5107
	N	5355	7145

Class	Precision	Recall
P	0.58	0.591
N	0.583	0.572

2) The Second list of attributes
Time taken to build model : 475.83 seconds
Accuracy : 64.248 %

		Predict	
		P	N
Actual	P	7229	5271
	N	3667	8833

Class	Precision	Recall
P	0.663	0.578
N	0.626	0.707

3) The third list of attributes
Time taken to build model : 431.95 seconds
Accuracy : 69.828 %

		Predict	
		P	N
Actual	P	8257	4243
	N	3300	9200

Class	Precision	Recall
P	0.714	0.661
N	0.684	0.736

3. Analysis and Observation

3.1 Significance of selecting 'attributes'

As it can be seen above, the first list has the lowest accuracy regardless of classifiers. Actually, the 10 selected attributes in the list does not contain any relevant information to distinguish if a given review is positive or negative. In the case of the second list, it mainly concentrates on 'positive' words which can possibly appear in positive reviews, which leads to higher accuracy by yielding a reasonable relevance with the goal of the classifiers. Lastly, the third list has the highest accuracy since it takes both 'positive' and 'negative' words into consideration. Taking care of the both sides has more advantages than only one side since it can prevent models from making biased decision.

Another interesting point is that even if a list of attributes consider both positive and negative words, if the choice of words is not plausible, the result still can be as poor as the case of the first list. For example, There is another list of attributes that consists of 270 words associated most with positive reviews and 69 words most with negative reviews, which were derived from the method mentioned above (Potts, 2011) , and they do not perform well on development data set. The accuracies are around 50% both for Naive Bayes and Random Forest. This proves that those words are not really appropriate to build useful models. In fact, these words are rarely used with low frequency over the collection so they scarcely give good standards. The second and the third lists are actually a compromise between selecting words by using certified technology (Potts, 2011) and by general common sense since the general 'good' and 'bad' words were selected just by their familiarity and intuition.

3.2 Differences between the classifiers

The most distinct difference between the two classifiers is 'time' taken to build models. As it can be easily observed, building models using Random Forest classifier requires much more time in comparison to Naive Bayes. The major reason is that in case of Random Forest, 100 random decision tress have to be built and as the number of attributes increases, it takes more time with more attribute options. However, in terms of accuracy of their models, they do not show any big difference. They yield similar outputs and sometimes, Naive Bayes produced a better model with higher accuracy. With this observation, we might be able to come to conclusion that the appearance of each word in the lists are not highly dependent with each other so that the 'naiveness' of Naive Bayes was able to demonstrate its competence.

3.3 How to improve models

There are a few ways to develop models and the most obvious one is to improve the list of 'attributes'. One of the possible ways to select useful words is to use human resources since it is 'people' who actually write reviews. For example, we can require a reasonable number of people to submit 50 positive and 50 negative words when they would use to write film reviews and the data can be mined to produce a list of attributes which probably are sorted by their frequency. This can give a good standard since it contains common sense from actual people.

4. Conclusions

In conclusion, the knowledge problem of deciding if a film is good or not relies on the ability to classify each review as being positive or negative toward its film and this problem is reasonably feasible if a good model can be built out of good selection of 'attributes'. Therefore, the crucial part is to conceive a convincing logic to sort out those attributes so that we can make the best use of the ability of classifiers.

References

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).

Potts, Christopher. 2011. On the negativity of negation. In Nan Li and David Lutz, eds., Proceedings of Semantics and Linguistic Theory 20, 636-659.