

Sentiment Classification for reviews based on Machine Learning

Author: Tianyi Ma

1. Introduction

With the development of Internet, we can always have a lot of data containing the sentiment of users which are called UGC. Usually we can use those to get the motion of the customer such as like or dislike of some products, agreeable or disagreeable of something.

We can use classification to get the reference of customer, therefore classification of the data by sentiment has widespread application. The goal of Sentiment Classification is looking for the emotion tendency of users by analyzing UGC. On account of Sentiment Classification is different from topic classification, not depending on topic words but understanding the emotion tendency of document. When I prepare to do this project I found out a lot research has been done in English so I decide to try finish sentiment classification on Chinese.

2. Relate work

First I go through the definition of Sentiment Classification and get the knowledge of it:

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service.

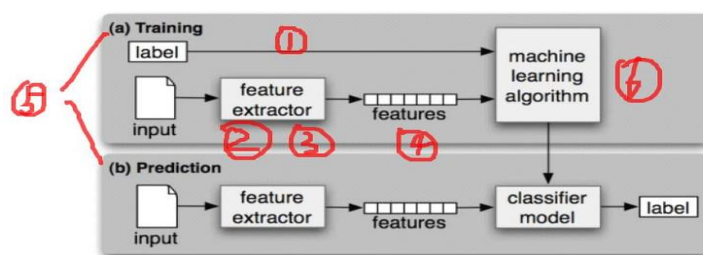


Figure 1 The construction of machine Learning

And I find some related work done by others the first is ‘Thumbs up? : sentiment classification using machine learning techniques’ which are finished by Pang B, Lee L, Vaithyanathan S. This helped me to go through the libraries of NTLKs

And the second one is Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums by Abbasi A, Chen H, Salem A., this one helps me to get a deep understanding of feature selection.

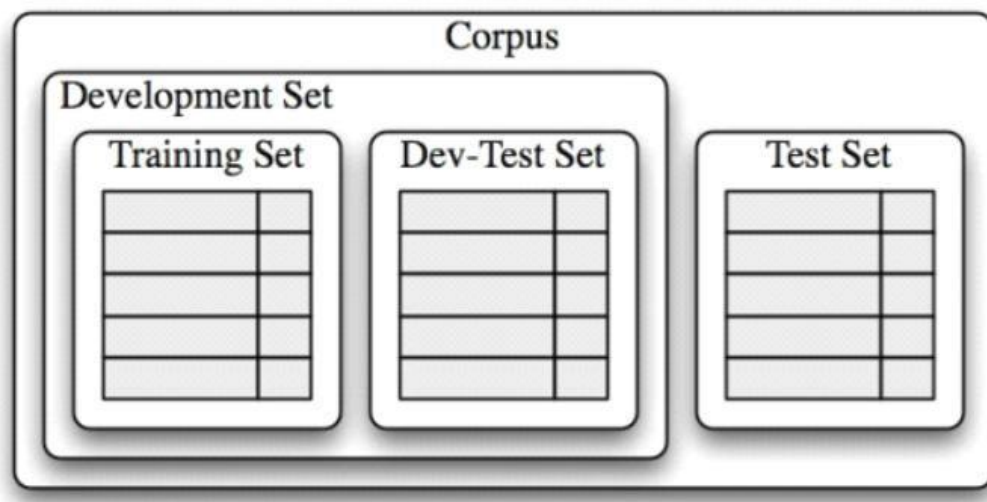


Figure 2 Train set & Test set

3. Data set

I want to learn more thing in this project so I decide to grab data by myself, the data I got is the comment of phone on website <http://mobile.zol.com.cn>, on account of I am not familiar with sql so I decide to use excel to store my own data.

First I learn how to use an application called Locoy Spider and use this to grab comments and some other element into .xls and then I use the library of python called xlrd to get the data into code for me to implement.

4. Proposed Solution & Experiments

Actually the analysis of emotional is to judge a sentence's emotion. It is passive or negative, on account of there are many examples in English, so I decide to grab some comments in Chinese to judge it.

4.1 Construction of project

The first one is data set. The second one is how to deal with those data there are some function of read txt file and excel file, doing Chinese word segmentation. I used library here, the first one is using xlrd and the second one is jieba to do Chinese word segmentation.

Next part is helping me to get the review helpfulness feature. I set:

6 features including review words and sentences number, review average length, review adjectives, adverbs and verbs number.

4 features including product name, brand and attributes' appearing times in a review. And review centroid score.

8 features including review positive/negative score, average score and standard deviation score. And review positive/negative probability score.

The last one is the most important part which use features calculated from above module as training set. I test five popular machine learning algorithm and use cross validation method to evaluate helpfulness prediction accuracy.

4.2 Algorithm of project

- 1 Read the comments and split them into single sentence.
- 2 Go through the emotional words and record the emotion and position of it
- 3 Search those words and calculate the score of it then save it into list
- 4 Calculate the average score and variance of each comments to get the conclusion.

The way to calculate score is just like:

“This phone is really good-looking and the function is so fancy, but the speed is so slow and the battery is bad, by the way the camera of it is so bad”

What I need to do is to figure out the word with emotion, such as “good”, “fancy”,

“bad”,when I meet something like good I will plus one on the score of sentiment, on the other hand I will minus 1 when I meet bad.And I also need to take care of something like “very”,

“so”, when I meet those word before or after emotion word above, I need to multiply the score above by 2. Such as “very good”: the score of it is $1*2$.

4.3 How to implement NLTK

I used NLTK in this project, when we have a split text we need to combine it into two or three words and find out the frequency of each word by using FreqDist, and next step is to find the words with information by using bigrams and we can set the data into an array to calculate the entropy and perplexity.

4.4 Test & Train data set

The pkl I used here is the pkl with positive comments and negative comments, what I have done is do the shuffle to the positive comments and let amounts of two comments same. And then give those comments feature label, I do it like this:

```
posWords = [feature_extraction_method(i),'pos']
```

And the last step is to split data into train and test,, and I do it like this:

```
train = posFeatures[174:]+negFeatures[174:]
```

```
test = posFeatures[:124]+negFeatures[:124]
```

So in the end I have 124 used as test and other used as train.

4.5 Classification and accuracy

On account I already have a test, so I need to split data and label, I use `zip(*)` here and then use `sklearn` in `nlTK`, then do `classifier.train(train)` and the next step is classify test and give the prediction and return the score, I used 3 class algorithm and `cv` to check the conclusion, there are: Naïve B, logistic regression, SVC, and then I used Chi-square to get the best information feature's amount.

And then I used the best conclusion to do the probability of positive and negative.

5. Results

	Bag of words	Bigrams	Best word feature	Best word bigram feature
Naïve B	0.79	0.71	0.87	0.91
LogisticRegression	0.81	0.75	0.73	0.8
SVC	0.71	0.79	0.77	0.8

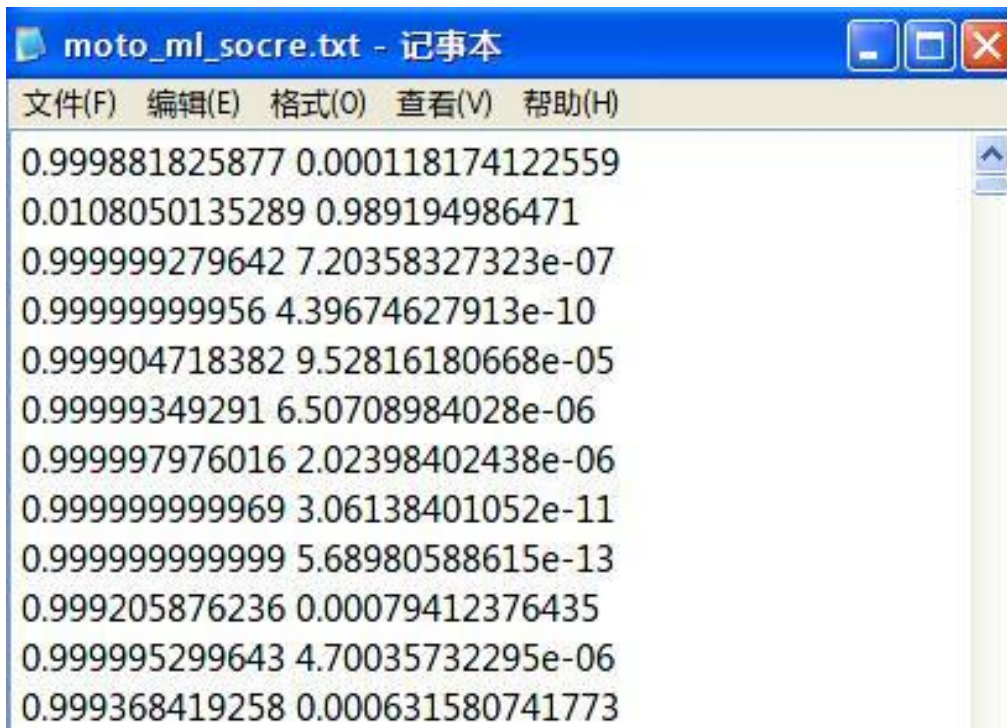
Table 1. The accuracy of each algorithm

The first one is normal data, the second one is I combine every two element data in data together and the third one is I set the value of information and choose the best 1500 data and the last one is I combine the word and choose the best 1500 of value of information (Feature Number).

	500	1000	1500	2000	2500	3000
Naïve B	0.88	0.86	0.87	0.87	0.85	0.85
Logistic Regression	0.85	0.85	0.86	0.85	0.83	0.83
SVC	0.81	0.76	0.79	0.7	0.8	0.8

Table 2. The accuracy with different Feature Number

I set naïve b and 1500 feature number to calculation the probability of positive and negative's value and the conclusion is like this:



0.999881825877	0.000118174122559
0.0108050135289	0.989194986471
0.999999279642	7.20358327323e-07
0.99999999956	4.39674627913e-10
0.999904718382	9.52816180668e-05
0.99999349291	6.50708984028e-06
0.999997976016	2.02398402438e-06
0.999999999969	3.06138401052e-11
0.999999999999	5.68980588615e-13
0.999205876236	0.00079412376435
0.999995299643	4.70035732295e-06
0.999368419258	0.000631580741773

Figure 3. probability of positive(left) and negative(right)

6. Conclusion

Thank you professor to let me do this project, and I also know how to grab data online and change them into the data I can use, if I have more time I think I can design a better algorithm to calculate the emotion score and it also can be used for company get the attribute of customers rapids, and I discuss with ruizheng he get me a lot of help in this project.

7.Reference

- [1] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.
- [2] Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification[C]//ACL. 2007, 7: 440-447.
- [3] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision[J]. CS224N Project Report, Stanford, 2009, 1: 12.
- [4] Kennedy A, Inkpen D. Sentiment classification of movie reviews using contextual valence shifters[J]. Computational intelligence, 2006, 22(2): 110-125.
- [5] Read J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification[C]//Proceedings of the ACL student research workshop. Association for Computational Linguistics, 2005: 43-48.