**Internet-based Information Extraction’s Project**

**--- Sentiment Analysis**

刘硕 5110309677  
齐轩 5110309683  
张智松 5110309684

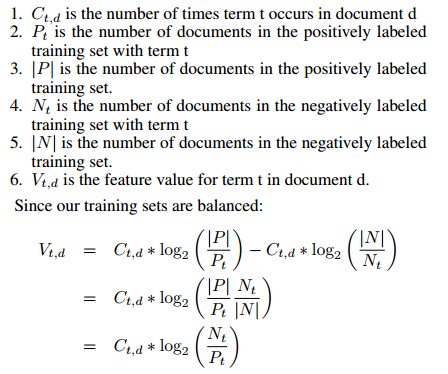
1. Introduction

Our task is document-level two-class sentiment classification. In this project, we focus on the task of classifying a document into two classes based on its sentiment orientation --- that is, it is positive or negative. This can be viewed as a special kind of document classification task, but it is kind of different because of its special goal of discovering the underlying sentiment of the document. And in this project, we focus on Chinese documents.  
 Maybe the most naïve thought is just count the so-called “sentiment word” in one document and use the counts to classify, which seems reasonable, however, we tried this method using the HowNet sentiment dictionary and get a result of only 59%. So we turn to machine learning methods.

We tried several methods for this task, and they are almost all supervised learning method, which is maybe the natural one if we have the training corpus. We use two classifiers, which are multinomial Naïve Bayes and SVM, and also several kinds of features with some feature selection method. Our final result is about 80% accuracy, maybe this is not really a high score, but we did improve much from the baseline method described below.

1. Baseline method

In fact, the central problem of the task is just feature engineering, which require us to extract more useful features that can represent the sentiment information. So the first thought is just the frequency of all words. We build a feature vector, which size is the dictionary’s length, and count the frequency of each word in one document. So, we represent one document with the word frequency counting vector, and feed it to MNB, and the result is about 72.1%, which we regard as our baseline.

1. Our trying  
   3.1 Suitable features   
    Several of our first trying are according a famous paper in 2002 by Pang & Lee [1], they use three supervised classifiers to analyze the sentiment and explore several features.  
    One finding is that use the existence of words or features are better than counting frequencies, and this leads to the bag-of-word features. And this method pushes our score up to 75.4% for MNB and 77.7% for SVM. Noting that the result is worse than their results for English movie data, which gets about 82%. So, we can see that this task really rely on language and sentiment corpus.  
    And they also explored several kinds of features such as using bigrams, unigram plus bigrams, adding part-of-speech information. We test several on our corpus and find that only bigrams features together with unigrams give the best result, and that is 77.7% for MNB and 78.9% for SVM.  
     
   3.2 Pre-processing  
    Another important idea is to pre-processed the document and ease the classification. We try 2 kinds: negation and although-but sentence, which are motivated by [1] and [2].  
    Negation is for sure one key annotation for sentiment, for it has the effect of turning the whole sentiment orientation to the opposite side. For dealing with negation, we use the simple method in [1], that is first find the key negation words, which in Chinese are just a few: "不","没有","不是","没","不如", and then add a prefix “\_NOT” to the words after those negation words in one sentence(up to a punctuation). This simple method works for MNB with 1% improvement and for SVM, it does not have clear effect.  
    Another observation is the conjunction words like “although” and “ but” in English. And for Chinese, we identify "虽说","固然","非但","虽然","尽管" for the meaning of Although, and "不过","但","但是","而是","反之","可是","然而","转而","恰恰相反","反倒","反而","却","仍" for But. We using the same trick of adding prefix of “ALTHOUGH\_","BUT\_”. And the results’ improvement is limited with just 0.2%, in fact this can’t be viewed as improvement, but we still think the “although-but” should be useful and we just do not deal with it well.  
     
   3.3 Feature selection  
    We also use a really simple feature selection method, which is motivated by [3], based on term frequency. We just filter out those features whose frequency either too high or too low in both positive and negative training examples. We set absolute boundaries and find this method gives a 1% improvement for MNB but some decrease for SVM.  
     
   3.4 delta tfidf  
    Our last trying is breaking away from the binary feature and return some “frequency” method, we use the method in [4], so-called delta tfidf.  
    The underlining principle is simple, we assign feature values for a document by calculating the difference of that word's TFIDF scores in the positive and negative training corpora.  
    The following graph from the origin paper illustrate the idea:  
     
   

Since now the features are no longer binary and we can’t use MNB, so we only test it using SVM and get the best result of 79.6%.

1. Results  
   Train and test the corpus on the course website (9:1 for train and test):  
     
   Accuracies (micro-averaged precision and recall):

|  |  |  |
| --- | --- | --- |
|  | MNB | SVM |
| (1)Unigram + negation | 76.6% | 77.5% |
| (2)Unigram + bigram | 77.8% | 78.9% |
| (3)Unigram + bigram + but | 77.9% | 79.1% |
| (4)Freq Feature selection | 78.5% | 77.6% |
| (5)delta tf-idf | Nope | 79.6% |

Precision, recall and f1 score:  
For MNB(multinomial naïve Bayes):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Negative | | | Positive | | |
|  | Precision | Recall | F1 | Precision | Recall | F1 |
| (1) | 75.6% | 77.8% | 76.7% | 77.8% | 75.6% | 76.6% |
| (2) | 76.0% | 80.3% | 78.1% | 79.7% | 75.3% | 77.4% |
| (3) | 76.0% | 80.5% | 78.2% | 79.9% | 75.3% | 77.6% |
| (4) | 78.1% | 78.4% | 78.2% | 78.9% | 78.5% | 78.7% |

For SVM:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Negative | | | Positive | | |
|  | Precision | Recall | F1 | Precision | Recall | F1 |
| (1) | 78.0% | 75.7% | 76.8% | 77.0% | 81.2% | 79.6% |
| (2) | 79.9% | 76.5% | 78.2% | 78.0% | 81.2% | 79.6% |
| (3) | 79.9% | 76.9% | 78.1% | 78.3% | 81.2% | 79.7% |
| (4) | 78.9% | 74.4% | 76.6% | 76.4% | 80.6% | 78.4% |
| (5) | 81.3% | 76.3% | 78.7% | 78.2% | 82.9% | 80.5% |

1. How to improve  
    In fact, maybe the central problem for using supervised learning for sentiment tasks is how to get good features, which must clearly notate the sentiment information of the socument.  
    In fact, our system is quite limited because it only uses the features of unigrams and bigrams, and we think that if we add syntactic information, we would get improved.  
    Another thought is the use of sentiment lexicons; in fact, we tried to use them but did not get any improvement. However, this must be really important features.
2. Implementations  
    Since we deal with Chinese documents, we have to do word-segmentation, we use NLPIR/ICTCLAS2014 [6] for this job, and at the same time it also gives us the pos information.  
    We also use the How-net sentiment dictionary [7], which is only used to help segment words in our system.
3. References  
   [1] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86, 2002.  
   [2] Fei Wang, Yunfang Wu, Likun Qiu: “Exploiting Discourse Relations for Sentiment Analysis”, Proceedings of COLING 2012  
   [3] Tan, S., & Zhang, J., “An empirical study of sentiment analysis for Chinese documents”, Expert Systems with Applications (2007)  
   [4] Justin Martineau, Tim Finin, “Delta TFIDF: An Improved Feature Space for Sentiment Analysis”, Third AAAI International Conference on Weblogs and Social Media, May 2009.  
   [5] Bing Liu, “Sentiment Analysis and Subjectivity” in Handbook of Natural Language Processing, Second Edition, 2010  
   [6] NLPIR/ICTCLAS2014 toolkit, http://ictclas.nlpir.org/  
   [7] How-net, http://www.keenage.com/html/c\_bulletin\_2007.htm  
   [8] Liblinear, http://www.csie.ntu.edu.tw/~cjlin/liblinear/  
   [9] A Practical Guide to Support Vector Classification, http://www.csie.ntu.edu.tw/