CS534 Machine Learning Project Proposal

——sentiment analysis on the amazon product

Xiao Tan, Haoyu Zhang, Kaiwen Zheng

Nov. 2017

1. Overview

We consider the problem of classifying users' review of products by sentiment, determining whether a review is positive or negative. It helps sellers to know users' like or dislike, and determine strategies for controlling numbers of products. Using the feedback of the products, it would help the salers, such as amazon to recommend product to customers. It also can help productors improve their products. To help potential customers get more information about products. The review sentiment analysis also can find the efficient data and evict some useless data such as some neutral data.

2. Background and related work

The basic idea to solve this kind of problem is Sentiment analysis, which refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to identify, extract, quantify and study affective states and subjective information. Now we focus on the part that write respect to some to topic or emotional reaction to a document, interaction, or even. Existing methods to analysis can be found in three main categories: knowledge-base techniques, statistical methods, and hybrid approaches.

As the previous work, the sentiment analysis would use the knowledge-base technology which would would base on the knowledge that would be error by human cognition. At the 2002, Pang B. and et al. publish their paper about the sentiment analysis using on the field of movie-review. They use the Navie Bayes, Maximum Entropy and SVM to find the classify model. They may find the different result by different exact model. They get about 83% at their best model. Alex Go and et al. achieve the sentiment anaylsis at the domin of Twitter. They also use the above three methods and use the different feature exactor to get the model dimension. Both two group would use the unigram, bigram feature as the dimension to get the result. In the unigram part, they get a similar result.

3. Techniques and Tools

As the field of exactor, we would use TF-IDF algorithm to get the feature of content. We would find a better way to get the model consider the knowledge-base techniques. Because the tf-idf would not consider the position of word, sometimes it would lead misunderstanding. We could use 2-3 classify methods from Naive Bayes, Maximum Entropy, SVM and other some methods. Some methods would be assessed by sklearn. Comparing the result and we would find a best model by them.

4. Data

Our dataset from http://jmcauley.ucsd.edu/data/amazon/. The data would be stored ass json as raw data. We need to exact the review data from a case then analyse each review and exact feature from each one. There is a massive review of amazon products. Due the size of the data, we can get a class product and split them as training set, dev set and test set. And these raw data include sorce 1-5 stars. Then we can pick 4,5 star review as positive label and 1,2 star review as negtitive label first.
If time is available, we could make the 3 star as neutral to find the result .

5. Evaluation

To evaluate the model, we will find the accuracy as the criterion. We would use the B. Pang's result as a criterion (about 80% in the accuracy). As for speed, due to the data size is large, we may choose a best size then make it run in a speed. And different methods has different time to create the model. We would also make sure let the training not run so long (fewer hours), but one hour training may be accepting.

6. Reference

[1] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.
[2] AlecGo, Richabhayani, LeiHuang. 2009. Twitter Sentiment Classification using Distant supervision [R]. Technical report, Stanford Digital Library Technologies Project.