Research proposal

Xiaoying Chen (s2714140)

06-12-2016

Introduction

Online shopping become more popular decent years. Customers buy products on the internet at wherever they are, and the products are sent to them. After the purchasing, people can leave their comments and ratings about the products on the products pages. This reviews represent the opinion of the customers about the products. In this research, I will analyze the online shopping reviews, and predict the rating that is given to the product by the reviewer base on the review with the most effective method.

Literature review

Turney (2002) developed an algorithm that classify the positive and negative reviews. The principle of the algorithm was to classify the reviews based on the average semantic orientation of the phrases that extracted from the reviews, which contained the adjectives (e.g., excellent and good) or adverbs. The average accuracy of using this method on reviews differed from every category. The average accuracy of 410 reviews from Epinions was 74%. The movie reviews were harder to classify, the accuracy was 66%. While the accuracy for automobiles and banks were about 80% to 84%. In the experiment of Hu and Liu (2004) indicated that using WordNet was an effective method to match the adjective semantic orientations in the reviews. The research of Dave et al. (2003) provided different approaches for feature extraction and scoring, an example feature extraction technique was divided the reviews into n-grams.

Material

For this research I will use the products reviews on Amazon. Since there are great number of products on amazon, and the corresponding review dataset are too large, I will only use the review dataset for the product category Office products for this research, the dataset comes McAuley et al. (2015). The information contains in the dataset are: reviewer ID, product ID, the name of the reviewer, the review, the average rating of the product, the summary of the review, and the time of the review. Duplicate reviews are removed from the dataset, and only the reviews for the products with more than 5 reviews are included, there are 53258 reviews in the dataset.

Methods

First, the reviews need to be tokenized, and the stopwords have to remove from bag of words. In order to extract the informative feature from the reviews, techniques of feature extraction will be needed. Some techniques that proposed in Hu and Liu (2004) will be using in this research. Use part-of-speech tagging to classify the type of each word (determine nouns, verbs etc.), to extract the opinion about the product (most time the opinion are adjectives, e.g., good, poor, and excellent). WordNet will be necessary to extract the adjectives in the reviews. Train the data by split the dataset randomly into training set (90%) and test set (10%). Classify the reviews with Nave Bayes classifier or Decision Tree classifier (choose the one with the higher accuracy). It is possible that more other effective methods will be applied to the program during the research.

Evaluation

The accuracy, precision, recall, and f-score have to be calculated. Perform the N-fold Cross-validation, calculate the average accuracy. Compare the average accuracy, and check if there is a significant difference with t-test. Consider the classification method with the highest accuracy as the most effective method.

References

Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM, 2015.

Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.