Research proposal

Xinzhe Yang

September 2017

Abstract

The yelp dataset provides a huge amount of customer reviews with number of ”useful”, ”stars” and ”funny”. It will be interesting to see what kind of reviews are more likely to be useful. If we can find an efficient algorithm to predict the usefulness or informativeness of review, Yelp can benefit by giving users writing suggestions, collapsing useless reviews (just like Quora collapsing answers) to improve user experience. It can certainly be applied to shopping sites, forums and any websites with reviews.

1 Introduction

Extensive research has been done on sentiment analysis for customer reviews. In a recent paper (1) using Yelp Dataset, the author applies the existing supervised learning algorithms and evaluates how well they predict star ratings based solely on text reviews. There are a lot more papers focusing on sentiment analysis. However, little attention is drawn to predicting the ”usefulness”.

In the paper (2), the author investigates a new method ”based on modeling the rate of learning of word meaning in Latent Semantic Analysis (LSA)” that estimates the informativeness of a term, which can be applied in information retrieval and summarization. However, the paper only investigates the informativeness for each word instead of the whole review sentence. But many ideas can be used in evaluating the sentence-level usefulness although ”usefulness” is not the same as ”informativeness” in the context of customer review.

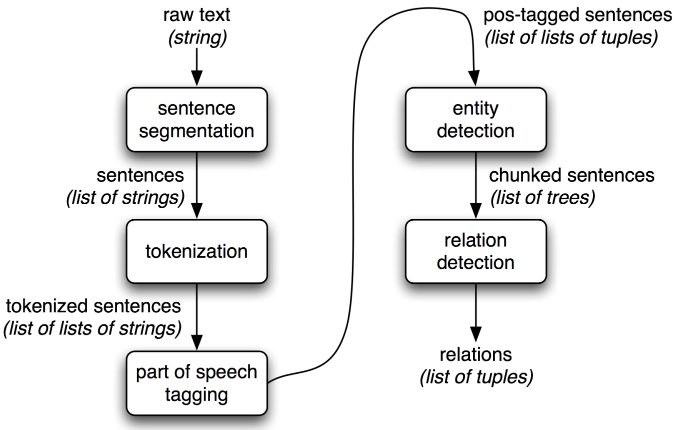
2 Data/Design

We will use the ”review.json” in Yelp Dataset. The file has 4736897 entries, each with text review, business id, date, user id, stars, and number of ”useful”’s. Sorting the useful column in descending order unsurprisingly shows that the top reviews are either the same restaurant or by the same user. Before using the data, we need to preprocess it by normalzing the ”useful” count corresponding to businessid and userid.

The goal of the research is to experiment different machine learning algorithms and develop evaluation metrics to see how to predict usefulness most accurately.

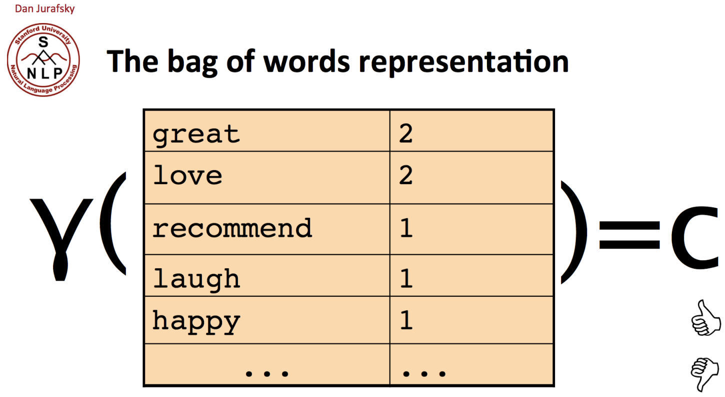
3 Methods

3.1 Feature Extraction



3.2 Naive Bayes

Fast, low storage
Irrelevant Features cancel each other without affecting results



Dan Jurafsky

The bag of words representation

$\gamma($

great	2
love	2
recommend	1
laugh	1
happy	1
...	...

$)=c$

👍
👎

3.3 SVM

`https://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf`
High dimensional input space

Few irrelevant features
Document vectors are sparse
Most text categorization problems are linearly separable

4 Significance

Yelp can benefit by giving users writing suggestions, collapsing useless reviews (just like Quora collapsing answers) to improve user experience. It can certainly be applied to shopping sites, forums and any websites with reviews.

5 Reference

1. Xu, Y., Wu, X., Wang, Q. (2015). Sentiment Analysis of Yelp‘s Ratings Based on Text Reviews.

2. Kireyev, K. (2009, May). Semantic-based estimation of term informativeness. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 530-538). Association for Computational Linguistics.

3. Nishikawa, H., Hasegawa, T., Matsuo, Y., Kikui, G. (2010, July). Optimizing informativeness and readability for sentiment summarization. In Proceedings of the ACL 2010 Conference Short Papers (pp. 325-330). Association for Computational Linguistics.