CS510: Advanced Information Retrieval
Project Proposal
Preference:Research Track

Qing Wang(Coordinator) Tianqi Wu
qwang55@illinois.edu twu38@illinois.edu

October 23, 2018

1 Introduction

Sentiment analysis is an important way to decide marketing strategy and improve product quality for e-commerce. However, there are challenges in generating accurate sementic anaysis such as ambiguous vocabulary and conjunction words. Our project implements sentiment anaysis on "Women E-Commerce Clothing Reviews" dataset from Kaggle for individual rating prediction. Our project also proposes a novel way of resolving conjunction words issue with respect to current methods.

2 Problem

- What is the research question?
- Why is this an interesting question to ask and why would we care about the answer to this question or a solution to the problem?
- Has any existing research work tried to answer the same or a similar question, and if so, what is still unknown?
- How to work out the answer to the question?
- How to evaluate solution?
- A rough timeline.

3 Proposed Approach

- The research question is: how to improve the accuray of sentiment analysis algorithm for e-commerce reviews based on conjunction words.
- It is an important qustion because the more accurate sentiment anaysis for e-commerce reviews is, the better guidance it provides on merchants' marketing strategy, product quality evaluation and customer service emphasis.
- There are related work that tries to address the question. Meena's paper [1] implements conjunction analysis which divides sentence into segments based on conjunction words and picks the best segment for sentiment analysis. Farooq's paper [3] focuses on negation anaysis which decides how to evaluate a sentence with negation words. Our project proposes a novel way of conjunction words analysis where instead of discarding the unimportant segments of a review, we fit sentiment analysis algorithm on each segment and then assign weights to each segment's rating to produce overall rating. We are still investigating on how to perform each step.
- To address the question, we first fit ordinal regression model or latent aspect model [2] on all reviews. Then we specify a set of conjunction words and divide each review into segments. Next we fit sementic analysis model on each segment and record their scores. Then we make a new dataset with segment scores for each conjunction word category. Then we fit this data with overall review ratings as labels to find weight for each segment. Finally we predict ratings for testing dataset based on the same rules and weights. We may change our procedures in practice if it improves our algorithm.

- To evaluate our results, we first split the raw data into training and testing sets. Then we fit our model on training sets and test on testing sets. We also plan to use cross-validation to reduce over-fitting and find best parameters for our model.
- Our temtative timeline:
  - Week 10: Finish preprocessing and algorithm design, including selecting conjunction words and divide reviews into segments.
  - Week 11: Finish algorithm selection for sementic anaysis. Finish fitting algorithm on original reviews and segmentized reviews and ratings recording.
  - Week 13: Finish algorithm selection for segment rating regression and finish up the codes for all model fitting parts. Finish progress presentation video.
  - Week 14-15: Tune parameters for best model performance and finish report.

References

[1] T. P. Arun Meena. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. 2007.

[2] C. Z. Hongning Wang, Yue Lu. Latent aspect rating analysis on review text data: A rating regression approach. 2010.

[3] A. N. Y. O. M. A. Q. Umar Farooq, Hasan Mansoor. Negation handling in sentiment analysis at sentence level. 2016.