Project Proposal

Name: Boxiong Tan

Date: 2015-3-29

Introduction

Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topics or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state, or the intended emotional communication.

A Research Question

This research intend to explore some methods in individual sentiment analysis.

Existing Solutions and Limitations

Existing approaches to sentiment analysis can be grouped into four main categories: keyword spotting, lexical affinity, statistical methods and concept-level techniques.

Each of these approaches has its limitations. The basic disadvantage of keyword spotting technique is the dependency on the presence of obvious affective words in the text. For instance, the emotion "sadness" cannot be derived from the sentence "I lost my money" as the does not specifically mention the word "sad".

Lexical affinity operating solely on the word-level can easily be tricked by sentences such as "I avoided an accident" (negation) and "I met my girlfriend by accident" (connotation of unplanned but lovely surprise).

In order to leverage statistical methods such as naive Bayes, support vector machine. One must first models the unstructured text into a computable representation, normally, in a form of vector or matrix. However, as the dictionary always contains a large number of vocabularies, the classifiers are easily affected by the curse of dimensionality.

My Solutions

I propose a method to improve sentiment analysis by utilizing the information from social network platform. The first information is comment. I do so for two reasons. First, comment information is easily obtainable. Second, more importantly, comments from friends always have hold similar attitudes. According to the principle of homophily [1], comment is one form of homophily that always shows similarity breeds connection. An study [3] found some evidence of homophily for both positive and negative sentiment among MySpace Friends.

The second information is previous content. Most sentiment analysis algorithms use simple terms to express sentiment. However, differing contexts make it extremely difficult to turn a string of written text int a simple positive or negative sentiment. Therefore, it is necessary to consider the last several tweets as an useful context.

Experiment Design

Data Collection

I planned to adopt the straightforward approach to creating a labeled test set, namely, extract an arbitrary user's tweets from Weibo (A Twitter like Chinese social network) or Twitter and label the tweets manually.

Modeling and Classification

In order to analyze the data and utilize a machine learning algorithm, the first step is to model the language. There are a few ways to model the language, bag of words for example, is a relatively easy approach. Trigram language model is also one of the most common way to model a language. In this project, I will explore both modeling techniques. Furthermore, I will develop classifiers on both models respectively and compare their results. The candidate classifiers include Naive Bayes, KNN and support vector machine.

Evaluation

Based on the label of each tweet, the result is evaluated with precision. However, according to research rater typically agree [2] of the time. Thus, a 70% accurate program is doing nearly as well as humans, even though such accuracy may not sound impressive. If a program were "right" 100% of the time, humans would still disagree with it about 20% of the time. Therefore, I planned to compare the performance with a popular sentiment analysis Web service: Google Prediction.

Expected Outcomes

Because of the methods that I used might not be the state of the art techniques, I expect the performance might not be as good as Google prediction. However, the performance should be higher than random selection, that is, higher than 50% accuracy.

References

[1] LAZARSFELD, P. F., AND MERTON, R. K. Friendship as a social process: A substantive and methodological analysis. In *Freedom and Control in Modern Society*, M. Berger, T. Abel, and C. Page, Eds. Van Nostrand, New York, 1954, pp. 18–66.

[2] OGNEVA, M. How companies can use sentiment analysis to improve their business. *N.p., n.d. Web* (2010).

[3] THELWALL, M. Emotion homophily in social network site messages. *First Monday 15*, 4 (2010).