Visual Tweets Sentimental Analysis

Lilong Jiang (jiang.573)

April 21, 2014

1 Introduction

In this project, I build a visual online tweets sentimental analysis tool, which allows users to do the sentimental analysis for a certain topic. The user interface is shown in Figure 1. We implement two classifiers and compare the performance and accuracy in both classifiers.

2 Dataset

We get our datset from the Sentiment140 [1]. The training dataset in this dataset is collected by the Twitter API, whose content contains the emoticons. The test dataset in this dataset is manually annotated.

3 Preprocessing

There are several challenges for the tweets processing: firstly, the tweets are very short, limited as 140 characters. Secondly, the tweets are very sloppy and it contains a lot of misspellings, slangs, abbreviations and emoticons. The preprocessing step is very important for the later classification. We preprocess all the tweets as follows:
(1) Remove usernames and urls.
(2) I use an emoticon dictionary [?], in which each emoticon is labeled as emotion(positive or negative). The emoticons in the tweets are replaced by happy or sad.
(3) NLTL[2] is used to tokenize the tweets, remove the stop words, punctuations, and lemmatization.
(4) PyEnchant[3] is used to check whether the token is a correct English word.

[1] http://www.sentiment140.com/

[2] http://www.nltk.org/

[3] http://pythonhosted.org/pyenchant/

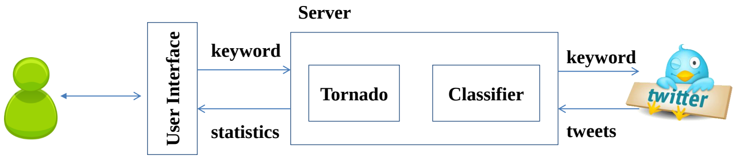

Figure 1: Architecture

4 Classifiers

We use two classifiers in this project. One is the naive Bayes classifier and the other is the SVM.

4.1 Naive Bayes Classifier

4.2 SVM Classifier

4.3 Experiment

We compare these two classifier on the performance and accuracy. 2000 tweets are used to train the classifers (1000 tweets are positive and 1000 tweets are negative) and 359 tweets are used to test the classifier.

5 Online Sentimental Analysis

5.1 Architecture

In the front end, we use d3[4] to visualize the final result and Tornado[5] as a webserver. When users submit a topic, the server gets 1000 tweets from the tweeter and analyzes each tweet and returns the result to the client.

5.2 Data Collection

5.3 Online Analysis

[4] http://d3js.org/

[5] http://www.tornadoweb.org/