Public Opinion Toward CSSTA: A Text Mining Approach

Yi-An Wu[*], Shu-Kai Hsieh[*]

Abstract

Recently, the public of Taiwan has had a heated debate on the issue of Cross-Strait Service Trade Agreement (CSSTA). After months of simmering tensions between ruling party and opposition party strongly backed by the student-led Sunflower Movement, the debate has finally reached a breaking point on March 18, 2014, at which students occupied the Legislative Yuan. During this period, novel communication such as Facebook sharing, instant messaging, and discussions on PTT have reshaped the social movement since they are easily accessible and instantly responded. The social media has become the dominant source in opinion shaping and the accompanying sentiment spread.

The extraction and tracking of uprising political opinions and events such as CSSTA has become one of the most important topics that receive much attention. With the huge amounts of texts, it is not possible to analyze and interpreting the social and political texts manually. Instead, we propose to use the text mining approach, which automatically extract opinion and information profiles from the texts. Moreover, this approach also strengthens the objectivity, for the norms are set *a priori*, and thus human biases are reduced.

As a pioneering work in the context of Taiwan society, this research aims to trace the public opinion toward CSSTA from the perspective of text mining. The approach involves the manually extracting of political stance related keywords and phrases, supervised machining learning, and a statistical model of the trend. We focus on the individual posts on PTT rather than news since they are more representative. The potential political or commercial applications are valuable. One can discover the public opinion and response in a short time.

The materials we used in this research includes a list of manually created seed words and phrases representing the pro-and-con political polarity, respectively. These words are tested by the texts of the website "服貿東西軍"[†], which classifies the supporting and opposing texts. Another resource we used in this work is the PTT corpus, which is a popular online bulletin board favored by many of the youth.

[*] Graduate Institute of Linguistics, National Taiwan University

[†] http://ecfa.speaking.tw/imho.php

Our procedures follow the common text mining techniques: features extractions, Chinese word segmentation with custom dictionary, establish the model for the SVM classifier, and using the N-fold cross validation for evaluations. We choose the keywords as the first step since many terms can potentially reveal one's attitude. For instance, the supporter for CSSTA would call students "霸佔", *occupy*, the parliament, while the opponent would use "留守", *stay*, in the parliament. Then we use these keywords as features to train the SVM classifier. The gold standards of the texts are chosen from the "服貿東西軍" website. The results are shown as follows:

Accuracy	Precision	Recall	F-score	Std. Dev.
0.850	0.850	0.859	0.855	0.040

We further extended our results to do the trend analysis. First, we apply the information gain calculated from the previous classifier, and then we sum keywords of each post, and sum over the posts of the same day. In other words, the score of each date is calculated as the following equation:

$$\text{Score} = \sum_i \sum_w IG(w) * C(w),\ i = \text{post index}, w = \text{word}$$

The figure demonstrates the popularity of this topic of each day. Second, we calculate the supporting information gain over the total information gain, and also sum over the posts in one day. This figure shows the ratio of supporting CSSTA from the analysis of posts.

Mining and tracking political opinions from texts in the social media is a young yet important research area with both scientific significance and social impact. The goal of this paper is to move one step forward in this area in Chinese context. We started from the manually created keywords and key phrases of CSSTA, used them to build a classifier and calculated their information gain, and then did the trend analysis of the PTT corpus. This approach involves interdisciplinary fields including information retrieval, data mining, statistics, machine learning, and computational linguistics. We hope that this text mining approach could discover the public opinion toward CSSTA, and further reveal political stances. Future works include more sophisticated language processing techniques applied to more broad domain of political topics, as well as developing dynamic tracking system gearing up for year-end election 2014.