Computational Linguistics and Chinese Language Processing
Vol. 19, No. 4, December 2014, pp. 19-28

Public Opinion Toward CSSTA: A Text Mining Approach

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

Abstract

Extracting policy positions from the texts of social media becomes an important technique since instant responses of political news from the public can be revealed, and also one can predict the electoral behavior from this information. The recent highly-debated Cross-Strait Service Trade Agreement (CSSTA) provides large amounts of texts, giving us an opportunity to test people's stance by the text mining method. We use the keywords of each position to do the binary classification of the texts and count the score of how positive or negative attitudes toward CSSTA. We further do the trend analysis to show how the supporting rate fluctuates according to the events. This approach saves human labor of the traditional content analysis and increases the objectivity of the judgement standard.

Keywords: Policy Position, Opinion Mining, Politics, Social Media, Trend Analysis

1. Introduction

Deriving reliable estimates of public opinions is central to the study of electoral behavior and policy positions. Among different methods, linguistic strategy has been one of the most widely used approaches in related studies in the field of political communication. For instance, budge *et al*. (1987) utilizes discourse-level opinion interpretation and stance recognition; while laver *et al*. (2003) and klemmensen *et al*. (2007) treated the words as “data”' encoding information about the political position of the texts' author. In addition to theoretical surveys, there are also numerous appealing applications on the political positions such as *abgeordnetenwatch*[1], where citizens are able to ask the members of parliament questions and express their attitudes through surveys, and the members of parliaments also respond to the questions. The dynamic design often attracts large organizations and political parties to keep a close eye on how the public form and represent their political stance, thus enhancing the

[*] Graduate Institute of Linguistics, National Taiwan University
E-mail:

[1] http://www.abgeordnetenwatch.de

transparency and accountability in the development of democracy.

Over the past few years, the production of huge volume of textual data has become an essential part of our current social life. In this context, there have been growing interests in applying text mining techniques to support Natural Language Processing applications in social and political domains, ranging from subjectivity and opinion mining, to ontologies and knowledge discovery. More and more attentions have been paid to the analysis and prediction tasks from the social media (Tumasjan *et al*., 2010; Conover *et al.*, 2011; Bermingham & Smeaton, 2011), which set a new scene for the data-driven research paradigm for social and political domains.

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. This action of “Occupy Taiwan Legislature” marked the beginning of a series of different political negotiations and efforts on this topic for both sides till April, 7. During this period, as well-recognized by many, using of novel communication technologies - Facebook sharing, instant messaging, sparking discussions on PTT, cloud documentation, etc - have reshaped the social movement not only domestically, but also globally.[2]

The uncertainty among members of society over the implementation of CSSTA is palpable. Due to its nature of easy access and instant response, 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 has thus become one of the most important topics that must be now be reckoned with. Though the task of analyzing and interpreting the social and political texts has gained its popularity in NLP-aided Social Science related fields, with the huge amounts of texts, it is not possible to analyze them manually. Instead, we propose to use the text mining approach, which automatically extract opinion and information profiles from the texts. In addition, this approach also strengthens the objectivity, for the norms are set *a priori*, thus human bias is reduced.

Our work is motivated by the compelling study of Junqué de Fortuny *et al*. (2012) which analyzed political opinions in Belgium by text mining of the newspapers. They used sentiment analysis to detect the opinion of the texts, and found the trends over timeline. Gelbukh *et al*. (1999) also used text mining techniques to analyze the Internet and newspaper news. They extracted the information of the texts by three steps: finding the topic of the document,

[2] Interested readers can refer to the cloud folder at http://hackfoldr.org/congressoccupied/ and the popular forum at http://www.reddit.com/r/IAmA/comments/21xsaz/we_are_students_that_have_taken_over_taiwans

extracting the opinion paragraphs by pattern matching, and matching topics with opinion paragraphs. They intended to discover how society interests are changing and to identify important current topics of opinion.

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.

This paper is organized as follows: first, we introduce some backgrounds of the studies of policy positions in section 2. Our approach to this topic and also the materials we used is described in section 3. The validity of our approach and the results are shown in section 4. Section 5 concludes the paper and suggests future works.

2. Previous Works

There is a growing body of studies on the topic of analysis policy positions. One traditional approach is content analysis, such as the Comparative Manifestos Project (CMP) (Budge *et al*., 1987; Benoit & Laver, 2007; Slapin & Proksch, 2008), where thousands of manifestos over 50 countries are interpreted by human decoders. However, this approach is so costly that it requires a huge amount of human labor. Another approach is computerized coding schemes (Kleinnijenhuis & Pennings, 2001), which match the texts to coding dictionaries. Laver and Garry (2000) created a dictionary of policy position which contains the predefined categories of political issues and the corresponding words. However, the approach also require much human labor on building dictionaries, and the words are insensitive to the contexts.

A variant of the second approach is the research of Laver *et al.* (2003), where they compared words in two different types of texts. One is the reference texts whose policy positions are defined *a priori*, and the other is the virgin texts whose policy positions are unknown but need to be found out. This approach is similar to the conventional **keyness** calculation where the *salient* keywords in target texts are measured and weighted statistically in comparing with the reference texts. However, as mentioned in (Klemmensen *et al*., 2007), the validity of the positions obtained by the this approach is “dependent on the choice of reference text and the quality of the a priori scores attached to these reference texts.” This poses a challenge for us because of the lack of representative reference corpus that can reflect the current language usage.[3] In this study, we adopt the second approach with a little variation, i.e. we also built the dictionary and tested its validity. More detailed procedures are explained

[3] Note that Sinica Corpus had ceased to update around 17 years ago.

in the next section.

3. Methodology

3.1 Materials

The material we used in this experiment includes a list of manually created seed words and phrases representing the pro-and-con political polarity, respectively. 8 linguistic graduate students from NTU were asked to compile the list based on their observations on the texts with CSSTA debate. It is noted that the keywords may be a word, a phrase, or a sentence. After some preprocessing, there are in total 350 terms for supporting CSSTA and also 350 terms for opposing CSSTA. We also use the texts on the website "服貿東西軍"[4] to be our gold standards of supporting and opposing texts. The selected texts are used to do the evaluations of our keywords.

Another resource we used in this work is the PTT corpus, a social corpus which has been constructed and dynamically updated by LOPE lab at National Taiwan University[5]. As an online bulletin board favored by many of the youth, PPT is doubtless the largest public forum and social media in Taiwan, with more than 1.5 million registered users and over 150,000 users online during peak hours. Many newest information are posted instantly on the Gossiping board. We analyzed every post on Gossiping board from January 1, 2014 to July 1, 2014, in total around 150,000 posts.

3.2 Procedures

Basically, we follow the text mining techniques suggested by Gupta Gupta and Lehal (2009), e.g. feature extraction, search and retrieval, categorization, and summarization. The detailed procedures are described as follows.

- Extract features.

  We arranged the works of every person with the unified format, which includes the keywords and the corresponding texts. Then we save the data in CSV files.

- Open-sourced Chinese word segmentation with custom dictionary.

  In order to flexibly fit the target texts, we extend an open-sourced Chinese word segmentation system[6]. There are many long keywords in the texts, which needs to be reserved in segmentation, so we first create the user dictionary of every keyword and load it to Jieba before word segmentation.

[4] http://ecfa.speaking.tw/imho.php

[5] http://140.112.147.131/PTT/

[6] https://github.com/amigcamel/Jseg

- Establish the model for the classifier.

After segmentation, each text is saved as an document (a vector of features and weights). The weighting scheme of the model is TFIDF and the classifier is a SVM classifier, which separates the documents in a high-dimensional space by hyperplanes.

- Use cross validation for evaluations.

N-fold cross-validation performs N tests on a given classifier, each time partitioning the given dataset into different subsets for training and testing. The indices for evaluations are accuracy, precision, recall, F1, and standard deviation.

- Calculate the information gain from the classification model.

Information gain is a measure of a feature's predictability for a class label. Some features occur more frequently with definite type of texts, so they are more informative. The information gain is defined as

$$IG(T,a) = H(T) - H(T|a),$$

where H is the Information Entropy

$$H(X) = -\sum_{i} P(X_i) \log_2 P(X_i)$$

The information gain is the entropy reduced by adding the new feature *a*.

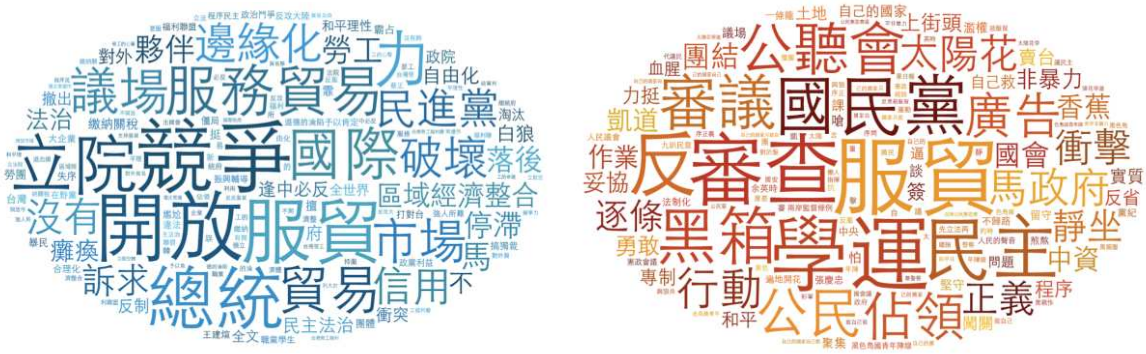
- Use the information gain to evaluate the texts from the PTT corpus.

We search the keywords of every post. Each keyword has the weight of the information gain. We sum over the information gain to judge the stance of the post, and then the scores of every post are further summed up in a day in order to observe the daily trend.

4. Results and Discussion

4.1 Keywords

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”, *occup*y, the parliament, while the opponent would use “留守”, *stay*, in the parliament. The supporter emphasized “信用”, *credibility*, “經濟”, *economy*, and “秩序”, *social order*, while the opponent would stress the “黑箱”, *black box*, “行動”, *action*, and “正義”, *justice*. The following are the word clouds for two types of keywords.



(a) Supporting keywords. (b) Opposing keywords.

Figure 1. Word clouds for supporting and opposing keywrods.

It is worth noting here that although opinion mining and sentiment analysis are often considered synonymous in many studies, it is necessary to draw the line between these two concepts. Following (Xu & Li, 2013), opinion is “a statement of the personal position or beliefs regarding an event, an object, or a subject (opinion target), while sentiment is the author's emotional state that may be caused by an event, an object, or a subject (sentiment target)”. So as reflected in the lists of keywords, we may find words representing certain opinions may be associated with a sentiment (e.g., “破壞”, *destroy*), but there are cases with standalone opinions (e.g., “開放”, *open*).

4.2 Classifier

We use these keywords as features to train the classifier. The gold standards of the texts are chosen from the “服貿東西軍” website. The cross-validation yields the results in the table 1.

Table 1. Cross-validation tests for the classifier.

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

There are about one third of keywords which can be found in our testing data. (supporting keywords: 116/350, opposing keywords: 136/350) The results show that 85 percent of the texts can be correctly classified as positive or negative opinion toward CSSTA by these keywords. Therefore, with the validity of our keywords selection, we are able to use the information gain of keywords to do the trend analysis.

4.3 Information Gain

From the classification model, we also obtain the information gain of each keyword. The information gain means to what degree the keyword contains the political polarity. The larger the information gain of a word, the greater probability of distinguishing two types of texts by the word. Some samples are shown in table2. Some keywords can distinguish the texts better like “競爭”, *competition*, and “反服貿”, *anti-CSSTA*, and thus they have more weights in classifying the texts.

Table 2. Information gains for two types of keywords.

Type	Keyword	IG	Type	Keyword	IG
support	競爭(Competition)	0.1242	oppose	反服貿(Anti-CSSTA)	0.1013
support	總統(President)	0.0862	oppose	學運(Movement)	0.1013
support	邊緣化(Marginalization)	0.0628	oppose	國民黨(KMT)	0.0996
support	破壞(Destroy)	0.0603	oppose	審議(Deliberation)	0.0804
support	落後(Fall behind)	0.0444	oppose	民主(Democracy)	0.0638
support	貿易夥伴(Trading partners)	0.0412	oppose	跳針(Skipping)	0.0628
support	利大於弊(Good than harm)	0.0402	oppose	行動(Action)	0.0528

4.4 Trend Analysis

While sentiments are always polar, it is not always the case for opinions. So instead of aiming to do binary classification of political texts only, we turn to use the information gain to do the trend analysis. First, 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:

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

where $IG(w)$ denotes the information gain of a word w, $C(w)$ denotes the word count of w, and the summation first sum over the word w in a post, then sum over the post i in a day. The reason why we sum up the values of IG's is that since IG is the change in information entropy, we can add up the entropy changes to see the tendencies of a text in the topic of CSSTA. Higher IG value means closer relations to the topic. The results are shown in the Figure 2. The corresponding events are listed in the Table 3. The figure demonstrates the popularity of this topic of each day, and the top spike remarkably indicates that the discussion on CSSTA increases abruptly from March 18, which was the date that protesters occupied Taiwan Legislative chamber, to the March 23, which was the date that some protesters further occupied the Executive Yuan.

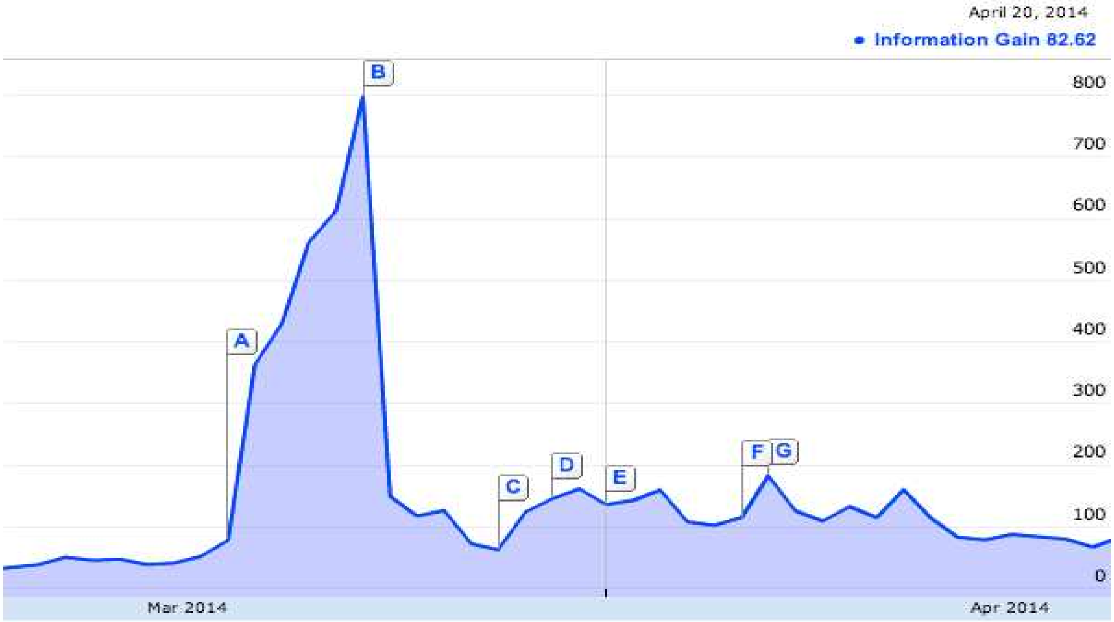


Figure 2. The trend of the topic popularity. (For the interactive figure, please click here.)

Table 3. Important events of Sunflower Student Movement.

Label	Date	Event
A	Mar. 18	Occupation of the Legislative Yuan
B	Mar. 23	Occupation of the Executive Yuan
C	Mar. 28	Rejection of the appeals by the Premier Jiang
D	Mar. 30	Demonstration
E	Apr. 1	March of the supporters
F	Apr. 6	Declaration of the President of the Legislative Yuan
G	Apr. 7	Announcement of the evacuation by the student leader Lin

The Figure 3 shows the ratio of supporting CSSTA from the analysis of posts. We calculate the supporting information gain over the total information gain, and also sum over the posts in one day. We can add the information gains like the previous analysis since the IG's are entropy changes. The information gains are added in both supporting and opposing aspects, and are compared to show the polarity of a text. The figure shows that the trend of supporting rate of CSSTA. The supporting rate drops on March 19, because of the Sunflower student movement. The supporting rate fluctuates for two possible reasons: the quantity of posts differs every day, and also the content of posts varies drastically. Thus the scores of the keywords varies in a wide range, which lead to the fluctuation of the supporting rate. But in general, we can see the tendency of the change.

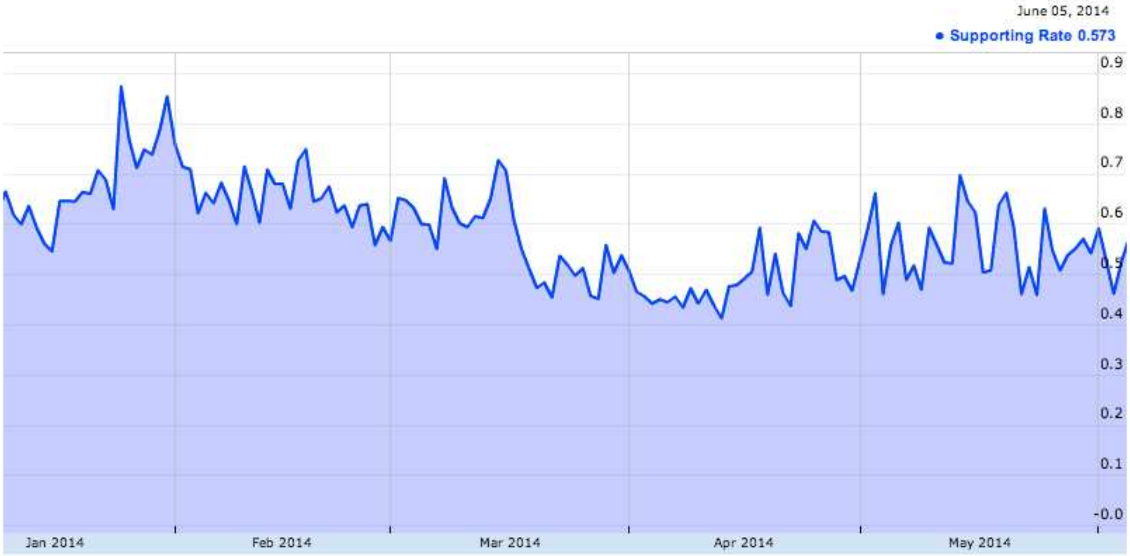


Figure 3. The trend of the supporting rate. (For the interactive figure, please click here.)

This method can be implemented on the coming election. The dynamic process of supporting rate for each candidate can be revealed by the texts on the social web, which is more efficient that the traditional telephone survey. Moreover, we can do more fine-grained analysis since the data is producing every day, and the   We can ask, for example, how the event or the speech of the candidates affect their supporting rate. There are huge potential of the political interests.

5. Conclusion

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.

References

Benoit, K., & Laver, M. (2007). Estimating party policy positions: Comparing expert surveys and hand-coded content analysis. *Electoral Studies*, *26*(1), 90-107.

Bermingham, A., & Smeaton, A. F. (2011). On using twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011*, 2-10.

Budge, I., Robertson, D., & Hearl, D. (1987). *Ideology, strategy and party change: spatial analyses of post-war election programmes in 19 democracies*. Cambridge University Press, 1987.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter." In ICWSM 2011, 89-96.

Gelbukh, E. F. , Gelbukh, E., Sidorov, G., Guzmán-arenas, A. *et al*. (1999). Text mining as a social thermometer," In *Proc. of the Workshop on Text Mining: Foundations, Techniques and Applications, IJCAI-99*. Citeseer, 1999.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence, 1(1), 60-76.

Junqué de Fortuny, E., De Smedt, T., Martens, D., & Daelemans, W. (2012). Media coverage in times of political crisis: A text mining approach. *Expert Systems with Applications, 39*(14), 11616-11622.

Kleinnijenhuis, J., & Pennings, P. (2001).11 measurement of party positions on the basis of party programmes, media coverage and voter perceptions. *Estimating the policy positions of political actors*, 2001, 162.

Klemmensen, R., Hobolt, S. B., & Hansen, M. E. (2007). Estimating policy positions using political texts: An evaluation of the wordscores approach. *Electoral Studies*, 26(4), 746-755.

Laver, M., & Garry, J. (2000). Estimating policy positions from political texts. *American Journal of Political Science*, 619-634.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(02), 311-3313.

Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3), 705-722.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10, 178-185.

Xu, G., & Li, L. (2013). *Social Media Mining and Social Network Analysis*. IGI Global, 2013.