Web mining for the mayoral election prediction in Taiwan

Introduction

Nowadays, many people express their opinions on the Internet, including opinions about political issues. Not only governments need to know about their citizens' opinions, but also most election campaigns and candidates are eager to know the voice of voters before the election so they can adjust their election strategy promptly. Over the years, election–related organizations have conducted polls forecasting, and the election results are mostly consistent with the prediction. However, in the Taiwan local elections held on 29 November 2014, the election results of many cities showed significant deviation from predictions. For example, the pre-election polls of the mayoral candidate of the Taoyuan city, Chih-Yang Wu, were 57 percent favorable, and those of his opponent, Wen-tsan Cheng, were only 25% favorable. However, the polls' laggard turned out to be the winner. Evidently, simply relying on the traditional polling method doesn't seem enough to provide an adequate forecast.

With the advent of social media, such as Facebook, Plurk and Twitter, users can express their opinions anytime and anywhere. In general, young people are the main users of the Internet. As shown in Table 1, 24% of the Internet access population in Taiwan are between the ages of 15 and 24, 29% of the Internet access population are between the ages of 25 and 34. Hence, using Web reviews for election estimation mainly reflects the opinions of young people.

****Insert Table 1 here****

There are many poll centers in Taiwan that generally collect data using traditional methods, such as via questionnaire and interview, which are expensive and time-consuming. In recent years, there has been increasing interest in the use of data mining to translate massive amounts of social media data into useful information. The new method of using Web mining for polling is not only fast and cost-saving, but also can be used to investigate the

opinions of populations that have not been reached before; it is therefore expected to replace or combine with traditional polling methods.

In view of the problems faced by traditional polling, such as inadequate data for quota design and poor response rates for phone polls, this study utilized the opinions posted on social media to gauge the degree of support for two mayoral candidates of Taichung city. However, to effectively and quickly extract useful information from tons of Web data is an important issue. In particular, there are many problems in conducting opinion analysis on comments in Chinese. Hence, before performing election prediction based on Chinese social media sites, the opinion extraction rules of Chinese texts should be established first.

The remainder of this paper is organized as follows. Section 2 introduces the development of social media mining in recent years with special focus on its application to election prediction. The problems of conducting semantic analysis on Chinese social media are also presented in this section. Section 3 depicts the framework of this study. The details of constructing opinion phrase extraction rules for identifying the opinion words associated with the attribute words are described. This section also introduces the six municipal governance-related topics. Section 4 presents the results of the scores of the six topics of each candidate, based on which the election prediction is performed. Finally, Section 5 presents conclusions and future work.

Related work

The practical study of public opinion is not new. Formal ways to measure public opinion are telephone surveys, focus groups, and content analysis. Since survey respondents occasionally provide socially acceptable answers rather than truthfully reporting their behavior, some researchers have devised research strategies that rely on unobtrusive analysis of remarks and speeches to evaluate decision makers' opinions (Ishiyama and Breuning, 2010). The unobtrusive research method does not require the cooperation of the subject under

investigation, and it is relatively low-cost compared with some traditional opinion survey methods. The increasing use of social media in recent years has provided public survey analysts with an opportunity to obtain a large amount of opinions without intruding in the research context. With the help of text mining-based information technologies, analysts can use a set of procedures to make valid inferences from reviews (Ampofo, et al, 2015). For example, Hodge and Matthews (2011) employed text mining to explore the possibilities of incorporating new media data in political environments.

The growing need for election forecasting has promoted the generation of much research using weblogs and social media to measure public opinion on elections. For example, Adamic and Glance (2005) analyzed blogs articles with respect to their political orientation, and constructed a network of links among political blogs in 2004 U.S. Presidential election. Albrecht et al. (2007) explored the use of campaign weblogs during the 2005 federal election in Germany including the distribution of blogs along party preference. Williams and Gulati (2008) have shown that the number of Facebook supporters can be regarded as a valid indicator of electoral success. In Wanner et al. (2009), a visualization is suggested to track the development of RSS news feeds over time that report on the U.S. presidential elections. Metaxas, et al. (2011) tested the predictive power of social media against several senate races of the US congressional elections. They pointed out that it is unlikely to guarantees electoral success simply because a candidate is scoring high in some social media metrics.

Among social media, Twitter data has particularly attracted significant attention. The use of Twitter data to forecast elections has become increasingly prominent although criticism has been levied against these studies for lacking methodological justification (Anstead and O'Loughlin, 2015). For example, Tumasjan et al. (2010) used Twitter to estimate the result of the 2009 German Federal election. O'Connor et al. (2010) advocated the potential of text streams to be used as a substitute and supplement for traditional polling

based on their findings of the close correlation between political opinions over the 2008 to 2009 period with sentiment word frequencies in Twitter messages. Ampofo, et al. (2011) analyzed Twitter content and the communication relationships in social media between citizens and political elites in the process of opinion formation in order to compare political elite tweets to citizen tweets. Sang and Bos (2012) emphasized that good election predictions cannot be obtained simply by counting Twitter messages mentioning political party names; instead, predictions can be considerably improved based on sentiment analysis and entity counts in tweets.

Bae et al. (2013) applied text-mining techniques to Twitter data related to the 2012 Korean presidential election and used a topic modeling technique to track changes in topical trends. Franch (2013) used an autoregressive integrated moving average (ARIMA) model to perform predictions for the vote share of the 2010 UK general election based on the political opinion of the audience aggregated from a number of social media platforms. Compared with the real vote share, Franch concluded that the proposed method well exceeded the predictive power of more traditional and expensive polls. Gayo-Avello (2013) performed prediction on election outcomes by incorporating Twitter sentiment into the computation rather than rely simply on volume. DiGrazia et al. (2013) used Twitter data to predict vote share for candidates in the 2012 U.S. Congressional elections. Ceron et al. (2014) used sentiment analysis to compute a Twitter popularity rating for Italian political leaders in 2011 and candidates in the French 2012 Presidential and legislative election. Smailović et al. (2015) proposed using an annotation platform and a binary support vector machine (SVM) classifier to real-time monitoring of the Twitter sentiment and showed its application to the Bulgarian parliamentary elections in 2013.

Overall, many studies appear to advocate that social media can serve as a useful tool to predict electoral outcomes across a variety of national contexts (Burnap, et al., 2016).

However, related research on Chinese data is rarely comprehensively studied. Several problems arise as we attempt to conduct semantic analysis on the Taiwan local elections of 2014: (1) Twitter is not popular in Taiwan. (2) Chinese opinions in natural language are generally expressed in more subtle and complex ways. (3) Since a Chinese sentence contains no delimiters to separate words, trying to find the possible word compositions of a sentence by comparing it with a lexicon may result in word segmentation ambiguities.

To address these problems, this study collects 4,419 data in total; among these, 2,009 data are collected from e-news (including Liberty times net, Times today, Chinatimes.com, UDN.com, ETtoday, TVBS), 167 data from magazines (Business weekly, Business today, Global views monthly, China times weekly, Storm media group), and 2,243 data from Facebook (Hu@thecharmingclub, Lin@forpeople, www.facebook.com/wethepeople.tc). We harvested these data from January 1 to November 28 in 2014, and employed technologies of text mining and Chinese semantic analysis to perform election prediction as detailed in Section 3.

Methodology

We disassemble the problem of opinion mining into the following subtasks: (1) Identify the municipal governance-related attribute words (2) Identify the opinions associated with attribute words (3) Determine the orientation/polarity of the opinions (4) Determine the degree/strength of the opinions (5) Determine the negative words (6) Score opinions based on their polarity and strength. Fig. 1 illustrates the framework of this study.

****Insert Figure 1 here****

After collecting data from the Web, we conduct word segmentation on the corpus and tag the part-of-speech (POS) of each word using the Chinese knowledge information processing group (CKIP), which is a Chinese parser developed by Academia Sinica. We select the attribute words manually, and then classify them into six municipal governance-related topics which are influential on public voting intentions. Next, we establish the opinion phrase extraction rules for identifying corresponding civilians' opinions associated with attribute words, and devise a scoring mechanism to transfer the reviews in each topic into scores. Based on the measure of the scores, we can not only predict the probability of being elected for each candidate, but also can identify which municipal governance-related political views or performance of each candidate are approved or detested by voters.

Opinion mining

Opinion analysis, a form of information extraction from text, has been of growing research and commercial interest in recent years, and this is no different for Chinese. Li et al. (2011) used the National Taiwan University Sentiment Dictionary (NTUSD), an emotion lexicon released by Taiwan University, to extract the orientation of features and used E-HowNet to determine their degree. E-HowNet is a lexical semantic representation model extended from HowNet to define lexical senses and achieve compositional semantics. In general, Chinese text is subtler and can lead to a higher degree of ambiguity than English text (Zhang et al., 2008). Among other things, Zhang et al. (2008) argued that Chinese opinion mining poses some challenges, including: (1) An additional step of segmentation is required. (2) The comparative and superlative sentences of Chinese using degree adverbs is much complicated than English sentences which mainly rely upon the suffix 'er' or 'est.' (3) Compared with English, Chinese opinion mining poses greater challenge for lack of adequate Natural Language Processing resources and corpus (Su et al., 2008). (4) Negative reviews

often contain many apparently positive phrases even if their authors maintain a strong negative tone (Zhang et al., 2008).

The municipal governance-related attribute words are words relevant to the civilians' concerns with the city government. Given attribute words, which are mostly nouns in a sentence, to establish the extraction rules for finding the associated opinion phrases is generally the following step to be performed. For example, Turney (2002) used a POS tagger to identify phrases in the corpus text. If the tags of two consecutive words conform to the proposed five patterns of tags, the matching adjective is identified as opinion, and matching nouns are extracted as object features. However, these patterns are oversimplified and suitable for English only. In their research to mine all the customer reviews of a product, Hu and Liu (2004) observed that an opinion word associated with an attribute word will usually occur in its vicinity; hence, they proposed to use the adjectives near the attribute words as the opinion words. However, Chinese text is subtler and can lead to a higher degree of ambiguity than English text (Zhang et al., 2008). Popescu and Etzioni (2007) used similar ideas to identify potential opinion phrases. Instead of using a window of size k, they took advantage of the syntactic dependencies computed by the MINIPAR parser and proposed 10 extraction rules for extracting the opinion words. Based on the study of Popescu and Etzioni (2007), the opinion phrases are no longer restricted to adjective only; they can be noun, verb or adverb phrases. Popescu and Etzioni (2007) argued that their precision on the feature extraction task is 22% better than most relevant previous review-mining systems.

Based on the above-mentioned approaches for identifying potential opinion phrases in English, this study aims to find rules for Chinese. After studying the grammatical structure of attribute words and opinion words in many reviews, we propose six patterns of opinion phrase extraction rules. The statistical results show that most of the opinion words are verbs, adverbs and nouns, and their relationships conform to the patterns in Table 2. In CKIP, each

POS tag includes several sub-class tags, for example, verbs are classified into transitive verb (VC/Vt) and intransitive verb (VA/Vi); adverbs are classified into quantity adverb (Daa), adverb before verb (Dfa), and adverb after verb (Dfb). In Table 2, N tags indicate nouns, T tags are auxiliary words, and attribute words are denoted as Na. The opinion words are usually near the attribute words. Thus, apart from establishing the extraction rules of opinion tags, the identification of the relationship between attribute words and opinion words is also required. The second pattern, for example, means that an opinion phrase is extracted if an attribute word is consecutively followed by an auxiliary word (T) and noun (N).

****Insert Table 2 here****

For the convenience of identifying the relationship between attribute words with opinion words, we use attribute words as centers and present the six patterns in Fig. 2. Some opinion words are in front of the attribute words, and some are after.

After identifying the opinion words associated with the attribute words, we compare the extracted opinion words with the NTUSD and determine their polarity/orientation. Where an extracted opinion word does not exist in the NTUSD, we manually include them in our opinion word database.

****Insert Figure 2 here****

The degree words are near adverbs to enhance the degree of opinion words. HowNet contains most of the popular emotion words and degree modifiers. This study uses HowNet to quantify the degree/strength of adverbs into five levels according to their intensities, including "超 (extreme) or 最 (most)", "很 (very)", "較 (more)", "稍 (over)", and "欠

(insufficiently)". We assign rating 6 to the opinion words having the degree level “extreme”, rating 5 for level “very”, rating 4 for level “more”, rating 3 for level “over”, rating 2 for level “insufficiently”, and rating 1 for a comment with no degree words. If a comment includes more than one degree word, we choose the one with the highest rating. For example, we assign rating 6 to this comment: "Chih-chiang Hu" + "實在 (really, rating 4) 厲害 (formidable)" + "也 (also) 超 (super, rating 6) 強 (powerful)".

Considering that negative words may reverse the polarity of the sentence, this study also establishes a negative words database, which includes 25 negative words as shown in Table 3.

****Insert Table 3 here****

The six municipal governance-related topics

To get a better understanding about the voice of voters, this study classifies the attribute words into six topics, including candidates' backgrounds (CB), transport infrastructure (TI), people's livelihood and public security (LP), social welfare policy (SW), arts and culture (AC), and local economy and industrial construction (LC).

The candidates' background relates to their political parties, previous or current positions, family members, achievements in their official career, political statements, and controversies due to personal opinions or actions.

The issues of transport infrastructure include all kinds of transport infrastructure to enhance traffic efficiency, including bus rapid transit (BRT), Light Rail Transit (LRT), free bus mileage, motorcycle parking spaces, elevated railway, ibikes, high speed railway, the No.74 expressway, taxi sharing schemes and rest stops, etc.

The issues of people's livelihood include food safety, environmental safety, traffic safety, disaster prevention, epidemic prevention, city appearance and council housing. The topics related to public security include social order, public safety, property and other lawful rights.

The issues of arts and culture include park and green land, night markets, department stores and business districts, art spaces, religion and ancient monuments, farm and tourism factory, eco parks, art and cultural activities, and cultural and creative parks.

Social welfare policy refers to social services provided by a government for its citizens. The related issues mentioned by the two candidates include: parenting, teenagers, youth, women, the elderly, socially vulnerable groups, low-income households, health insurance, and subsidies.

The issues of local economy include citizens' incomes, taxes, fees, unemployment, commodity prices, house prices, population density, import, and export trade. With respect to industrial construction, issues relate to Central Taiwan Science Park and industry parks, including the land utilization, transportation, electricity, water, communication, sewage and waste disposal.

The score calculation

We define the index of each attribute word as t_{ijk}, where i=1 refers to candidate Hu, and i=2 refers to Lin; k=1,…,6 refers to the topics of CB, TI, LP, SW, AC, and LC, respectively; and j refers to the attribute words that belong to each topic. The number of attribute words in each topic for each candidate maybe different. For example, in Figure 3, we show six major attribute words in the topic of CB for Hu: DPP, KMT, reelected consecutively, green party, political achievements, and blue party, respectively. As shown in Figure 4, there are five major attribute words in the topic of CB for Lin: DPP, Blue, green party, political views, and deep green. Hence, for instance, t_{132} refers to the attribute word "連任 (reelected

consecutively)", which belongs to the topic of CB of mayor Hu. We first assign value to the polarity of the opinion words by setting $P(t_{ijk})$=1 for a positive opinion and $P(t_{ijk})$= −1 for a negative opinion. Then, we assign values to degree words, $D(t_{ijk})$, which range from 1 to 6 according to their intensities as defined in the section of opinion mining. Hence, if an opinion word has a degree word nearby, the opinion word will have a score ranging from -6 to +6; the higher the value, the more positive the opinion word is. We define $N(t_{ijk})$ as the negation modifier, which is −1 or +1 denoting whether a negation word is present or not, respectively.

Similarly to the method of Popescu and Etzioni (2007), we transform each sentence into its corresponding quadruple $<t_{ijk}, P(t_{ijk}), D(t_{ijk}), N(t_{ijk})>$ and compute the score of topic k for candidate i by the following formula:

$$S_{ik} = \sum P(t_{ijk}) \times D(t_{ijk}) \times N(t_{ijk}) \times F(t_{ijk}) \quad (1)$$

where $F(t_{ijk})$ is the number of comments of attribute word t_{ijk}. The numbers of attributes words in each topic are different; therefore we use a summation symbol without index in equation (1). The total score of candidate i is computed as follows:

$$S_T(i) = \sum_{k=1}^{6} S_{ik} \times W_k \quad (2)$$

where W_k is the weighting of topic k, which is the percentage calculated by dividing the total number of comments of topic k (including the positive and negative comments of the two candidates) by the total number of comments, which is 1427. Hence, the weighting of the topics CB, TI, LP, SW, AC, and LC are 47.9%, 40%, 5.4%, 3.5%, 2.8% and 0.3%, respectively.

Data analysis and discussion

This study harvests election-related comments from e-news, magazines, and Facebook. During the process of extracting attribute words from the reviews, many synonyms were found, such as KMT being the equivalent of the Chinese Nationalist Party and DPP referring to the Democratic Progressive Party. We used phpMyAdmin to establish an attribute words database, and in total, 259 attribute words were included. Of these, 71 words belong to the topics of transport infrastructure, 44 words belong to arts and culture, 48 words belong to social welfare policies, 22 words belong to local economic and industrial construction, 33 words belong to people's livelihood and public security, and 41 words belong to candidates' backgrounds.

This study uses the opinion phrase extraction rules listed in Table 2 to identify the associated opinion words, which are then classified into two databases: the positive opinion words database and the negative opinion words database. Sentences with neutral opinion are not considered.

If there is a negative modifier near a positive opinion word, this sentence is classified as a negative comment. For example, "林佳龍 (Chia-Lung Lin)" + "不是 (is not, negative words)" + "很 (very, degree words)" + "適合 (suitable, opinion words)" +當市長 (as a mayor)". Conversely, if a negative modifier is near a negative opinion word, this sentence becomes a positive comment. For example, "林佳龍 (Chia-Lung Lin)" + "並非 (really isn't, negative words)" + "很 (very, degree words)" + "爛 (bad, negative opinion words) ". The sentence would maintain its orientation if no negative modifier was found.

The numbers of positive comments and negative comments about each topic are shown in Table 4, in which "+" and "−" denote positive and negative, respectively. Apparently, citizens show more concern on the issues of candidates' backgrounds and transport infrastructure. We use mayor Hu's positive score of the topic CB as an example to show how

the score is calculated. The Hu's numbers of positive comments of the topic CB in each degree level are shown in Table 4.

****Insert Table 4 here****

Hence, the positive score of mayor Hu (i=1) of topic CB (k=1) is calculated as $(1\times 22+2\times 54+3\times 14+4\times 44+5\times 28+6\times 70)\times 0.479=435$. Following this procedure, the scores of overall topics of the two candidates are shown in Table 5. For topic CB, mayor Hu has a positive score of 435 and negative score of 305.6 while Chia-Lung Lin has a positive score of 527.8 and a negative score of 385.2. Besides the scores, the number of comments, which refers to the discussion frequency of the voters, may also provide some interesting information. For example, the fact that both the positive and negative number of comments of Lin's are higher than that of Hu's shows that more people are talking about Lin. Based on the data of the percentage of degree words of the "extreme" level presented in Table 5, the comments on Hu tended to be divided into two opposite extremes, i.e., many people express their opinions (both positive and negative) on Hu with intense wording, while the comments about Lin are comparatively mild.

****Insert Table 5 here****

This study uses correspondence analysis to transform opinion information on the two candidates into graphical displays to facilitate the interpretation of voters' views. Being a useful tool to uncover the graphical relationships among categorical variables, correspondence analysis has been applied to a wide range of research areas including education, marketing, and image retrieval (Shanka, 2006). Correspondence analysis is

performed on a contingency table, which is in a matrix format that displays the frequency distribution of the variables. By decomposing the chi-squared statistic associated with this table into orthogonal factors, this study displays a set of data of positive and negative evaluation of the two candidates in two-dimensional graphical form using software XLSTAT.

In Taiwan, the representative color of KMT is blue, and that of DPP is green; i.e., "blue party" refers to KMT and "green party" refers to DPP. As shown in Fig. 3, for the candidate of the blue party, mayor Hu, the attribute word "國民黨 (KMT)" was mentioned 161 times as a positive comment and 160 times as a negative comment, meaning that people for and against Hu because of his party label are about the same quantity. A similar case happened to Chia-Lung Lin. The most often mentioned attribute word about Lin was "綠黨 (green party)", which included 280 positives and 297 negatives, and the next one was "民進黨 (DPP)" which included 220 positives and 185 negatives. The numbers of positive and negative comments about the two candidates were about the same, meaning that many peoples already have preconceived views about the party and that the number of supporters of the two parties is approximately equal.

****Insert Figure 3 here****

It is worth noting that the attribute word "連任 (reelected consecutively)" was frequently mentioned since Hu served as Taichung mayor for eight years before being elected to his current position following the former city's upgrade to a special municipality. This issue received widespread attention in both parties. That the number of negative comments (74) is higher than that of positive comments (59) on this issue shows that although some people supported Hu's bid for re-election, more people opposed it.

Regarding the topic of "transport infrastructure" shown in Table 5, mayor Hu has a positive score of 137.6 and negative score of 187.2, and Lin has a positive score of 39.2 and negative score of 65.6. Since Hu is the incumbent mayor, there are more people talking about him and so both the positive score and negative score of Hu's are higher than the scores of Lin's.

****Insert Figure 4 here****

For mayor Hu, the attribute word "BRT" was most often mentioned with 121 positive and 196 negative comments. Moreover, he has 87 positive comments and 63 negative comments of the attribute word "路線規劃 (route planning)" and 25 positive and 13 negative comments of "輕軌捷運 LRT". BRT was one of the important achievements that Hu's campaign focused on promoting. The operational situation of BRT was of great concern to Hu's election success. The perceptual diagram in Fig. 5 is helpful for the explanation of the evaluation results. For example, BRT is nearer to the negative than the positive as shown in Fig. 5, meaning that BRT attracted more criticism than praise. It is noteworthy that the "route planning" has more positive comments than negative comments, but it positions farther into negative than positive as shown in Fig. 5. This may imply that Hu has received strong negative comments due to his ignorance to the transport need of people living in the suburbs.

****Insert Figure 5 here****

In contrast, as shown in Fig.6, Lin received 35 positive comments and 60 negative comments on his political views regarding route planning. On the whole, the number of negative comments is greater than that of positive comments for both candidates, showing

that people had high anticipation on issues related to transport infrastructure. Although Hu has a higher positive score than Lin, his negative score is even higher, indicating that the effect of his incumbency burden is larger than his incumbency advantage.

****Insert Figure 6 here****

Compared to the above two topics, other topics received much less attention from the Taichung civilians. For the topic "livelihood and public security" shown in Table 5, Hu has a positive score of 1.8 and negative score of 1.0. The results indicate that people are concerned about the problems of sinister gangs and food safety, and Hu has passable positive performance in these respects. For the topic "social welfare policies", Hu has a higher positive score than negative score, showing people are satisfied with Hu's subsidy policy. For the topic "Arts and culture", the term "歌劇院 (Opera house)" as shown in Fig.7 was located in the middle of the positive and negative, meaning the newly opened opera house brought Hu both positive and negative responses. Apart from the opera house, people highly valued the "花博 (flower exposition)", which has been held for years. Since Lin does not have any governance performance, comments related to that topic were rare.

****Insert Figure 7 here****

Summing the scores for each topic, the total scores for Hu and Lin are 81 and 116, respectively. The proportions of the scores of the two candidates are shown in Fig. 8. Based on this statistic, Lin outscores Hu by 17.74%. This outcome is close to the real election results in which Lin acquired 14.12% more votes than Hu.

****Insert Figure 8 here****

On the whole, among the six topics, "candidates' backgrounds" and "transport infrastructure" received the most attention from civilians; hence, the scores of these two topics dominated the election results. Although Hu had advantages over Lin in the remaining four topics, these had only little effect on the election results because the volume of discussions on these topics was small. In the most discussed topic, "candidates' backgrounds", both the positive score and negative score of Lin's are higher than that of Hu's. By subtracting the negative score from the positive score, Lin has a net score of 142.8, which is higher than Hu's net score 129.4. Therefore, the overall support that Lin received on the topic of "candidates' backgrounds" is better than Hu. This may be because the KMT party has held the reins of the city government for a long time, which results in people's anticipation for political rotation. On the topic of "transport infrastructure", both the positive score and negative score of Lin's are lower than that of Hu's. By subtracting the negative score from the positive score, Lin has a net score of -26.4and Hu has -49.6. In other words, there were more negative comments than positive comments for both candidates, meaning civilians are demanding a higher level of transport infrastructure. Hu was originally hoping that he could win points for his victory through the promotion of transport infrastructure; however, in the end, Hu lost more points than Lin in this topic, which turned out to be one of the major causes for his loss in the election.

Conclusion

In Taiwan, most election polls have been conducted using telephone calls for years. With the rise in popularity of the Internet and the rapid development of IT technologies, many people express their opinions about the election on Internet platforms. The major

contributions of this study are two-fold. First, in order to address the problem of the Chinese opinions, which are generally expressed in more subtle and complex ways, this study proposes new rules for the extraction of Chinese opinion words associated with attribute words. Second, this study applies Chinese semantic analysis to assist in predicting election results and investigating the topics of concern to voters. To analyze the opinions of the municipal governance-related issues of concern to the citizens in Taichung city, this study classified the collected comments into six topics, including candidates' backgrounds, transport infrastructure, livelihood and public security, arts and culture, social welfare policies, and local economy and industrial construction. Based on the scores of each topic, political parties and candidates can get a clearer idea about the public needs, and utilize this information to plan better policies. We used correspondence analysis to visually present the competition strength of the two mayoral candidates. That the prediction had only 3.62% difference from the real election results shows the effectiveness and accuracy of the proposed prediction method.

This study has some limitations that require further consideration in the future. First, this study focused on revealing Web opinions, and the poll results obtained from traditional method, such as telephone surveys and questionnaires, were not included. To get a more accurate election prediction result, merged methods should be conducted in the future. Second, the opinion database should be updated by considering some newly appeared symbols, such as "囧", an emoticon meaning embarrassment, and "XD", which indicates smiling with eyes narrowed. Third, the extraction rule should capture the ironic sentiments. For example, according to the extraction rule in this study, "○○○候選人好棒棒" would be classified as a positive statement since it includes "好 (very)" and "棒 (excellent)"; However, it is actually a negative statement. Fourth, a sentence that expresses an opinion with a question mark is usually presenting a negative statement or a weak positive statement.

However, since the pronunciations, including the question mark, are deleted via CKIP during the process of word segmentation, it may result in a missed interpretation. To support more accurate analysis of the public opinion evaluation with this type of interrogative sentence, further grammatical structure studies are required.

References

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 US election: divided they blog. In Proceedings of the 3rd international workshop on Link discovery (pp. 36-43). ACM.

Albrecht, S., Lübcke, M., & Hartig-Perschke, R. (2007). Weblog campaigning in the German Bundestag election 2005. Social Science Computer Review, Vol.25 No. 4, pp.504-520.

Ampofo, L., Anstead, N., & O'Loughlin, B. (2011), "Trust, confidence, and credibility: Citizen responses on twitter to opinion polls during the 2010 UK general election", *Information, Communication & Society*, Vol.14 No. 6, pp.850-871.

Ampofo, L., Collister, S., O'Loughlin, B., & Chadwick, A. (2015), "Text Mining and Social Media: When Quantitative Meets Qualitative and Software Meets People", *Innovations in Digital Research Methods*, Sage Publications Ltd, New York, NY, pp.161-91.

Anstead, N., & O'Loughlin, B. (2015), "Social media analysis and public opinion: The 2010 UK general election", *Journal of Computer-Mediated Communication*, Vol.20 No.2, pp.204-220.

Bae, J. H., Son, J. E., & Song, M. (2013), "Analysis of twitter for 2012 South Korea presidential election by text mining techniques", *Journal of Intelligence and Information Systems*, Vo.19 No.3, pp.141-156.

Burnap, P., Gibson, R., Sloan, L., Southern, R., & Williams, M. (2016), "140 characters to victory?: Using Twitter to predict the UK 2015 General Election", *Electoral Studies*, Vol.41, pp.230-233.

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014), "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and Franc", *New Media & Society*, Vol.16 No.2, pp.340-358.

DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013), "More tweets, more votes: Social media as a quantitative indicator of political behavior", *PloS one*, Vol.8 No.11, e79449.

Franch, F. (2013), "(Wisdom of the Crowds) 2: 2010 UK election prediction with social media", *Journal of Information Technology & Politics*, Vol.10 No.1, pp.57–71.

Gayo-Avello, D. (2013), "A meta-analysis of state-of-the-art electoral prediction from Twitter data," *Social Science Computer Review*, Vol.31 No.6, pp.649-679.

Hodge, B., & Matthews, I. (2011), "New media for old bottles: Linear thinking and the 2010 Australian election", *Communication, Politics & Culture*, Vol.44 No.2, pp.95-111.

Hu, M., & Liu, B. (2004), "Mining and summarizing customer reviews", in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining in Seattle, WA, 2004,* Association for Computing Machinery (ACM), New York, NY, pp. 168-177.

Ishiyama, J. T., & Breuning, M. (Eds.). (2010), "21st century political science: a reference handbook", Sage Publications Ltd, New York, NY.

Li, C. R., Yu, C. H., & Chen, H. H. (2011). "Predicting the semantic orientation of terms in E-HowNet", in *Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing, in Taipei, Taiwan, 2011*, Association for Computational Linguistics, pp.151-165.

Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011, October). How (not) to predict elections. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 165-171). IEEE.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010), "From tweets to polls: Linking text sentiment to public opinion time series", In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM), in Washington, D.C. USA, 2010*, the Association for the Advancement of Artificial Intelligence (AAAI), California USA, pp.122-129.

Popescu, A. M., & Etzioni, O. (2007). "Extracting product features and opinions from reviews", *in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing in Vancouver, Canada, 2005*. Association for Computational Linguistics Stroudsburg, PA, USA, pp. 339-346.

Sang, E. T. K., & Bos, J. (2012), "Predicting the 2011 Dutch senate election results with Twitter", in *Proceedings of the Workshop on Semantic Analysis in Social Media in Avignon, France, 2012*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp.53-60.

Shanka, T., Quintal, V., & Taylor, R. (2006). "Factors influencing international students' choice of an education destination–A correspondence analysis", *Journal of Marketing for Higher Education*, Vol.15 No.2, pp.31-46.

Smailović, J., Kranjc, J., Grčar, M., Žnidaršič, M., & Mozetič, I. (2015). "Monitoring the Twitter sentiment during the Bulgarian elections", *in IEEE International Conference on Data Science and Advanced Analytics (DSAA) in Paris, France, 2015*, IEEE, Piscataway, NJ, pp. 1-10.

Su, Q., Xu, X., Guo, H., Guo, Z., Wu, X., Zhang, X., Swen, B. and Su, Z. (2008, April). Hidden sentiment association in Chinese web opinion mining. In Proceedings of the 17th international conference on World Wide Web (pp. 959-968). ACM.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). "Election forecasts with Twitter: How 140 characters reflect the political landscape", *Social Science Computer Review*, Vol. 29 No.4, pp.402–418.

Turney, P. D. (2002). "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews", *in Proceedings of the 40th annual meeting on association for computational linguistics in Philadelphia, PA, USA, 2002*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 417-424.

Wanner, F., Rohrdantz, C., Mansmann, F., Oelke, D., & Keim, D. A., "Visual sentiment analysis of RSS news feeds featuring the US presidential election in 2008", *in Visual Interfaces to the Social and the Semantic Web (VISSW 2009) in Sanibel Island, Florida, USA, 2009*, Association for Computing Machinery (ACM), New York, NY, USA, pp.1-8.

Williams, C., & Gulati, G. (2008). What is a social network worth? Facebook and vote share in the 2008 presidential primaries. In Annual Meeting of the American Political Science Association, 1-17. Boston, MA.

Zhang, C., Zeng, D., Xu, Q., Xin, X., Mao, W., & Wang, F. Y. (2008). "Polarity classification of public health opinions in Chinese", *in International Conference on Intelligence and Security Informatics in Taipei, Taiwan, 2008*. Springer, Berlin Heidelberg, pp.449-454.

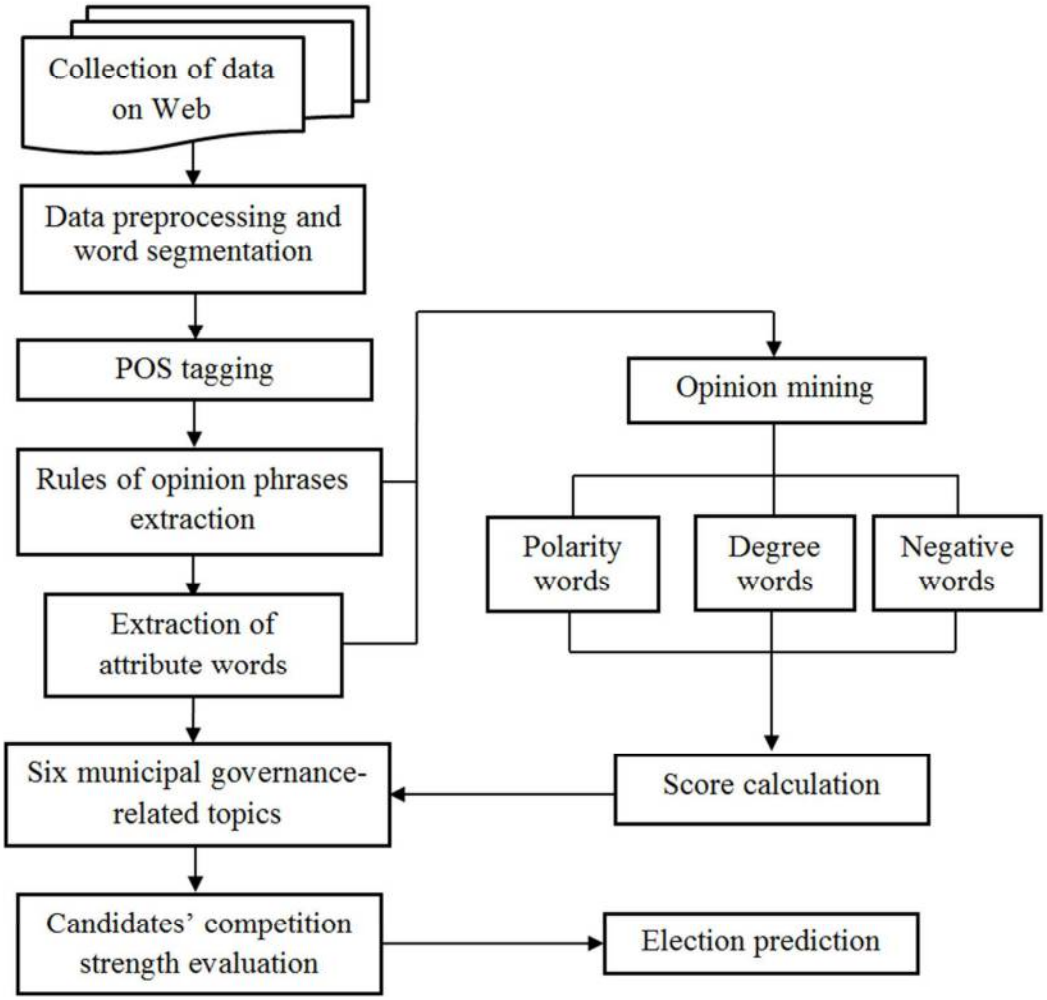

Figure 1 Framework of this study.

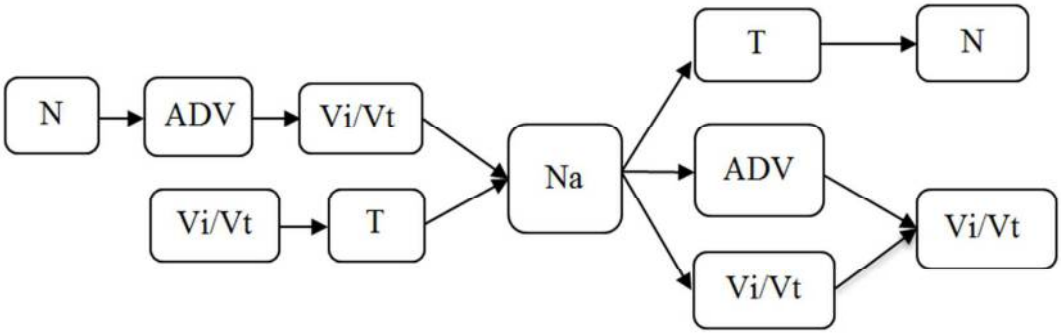

Figure 2 Six patterns of opinion phrase extraction rules.

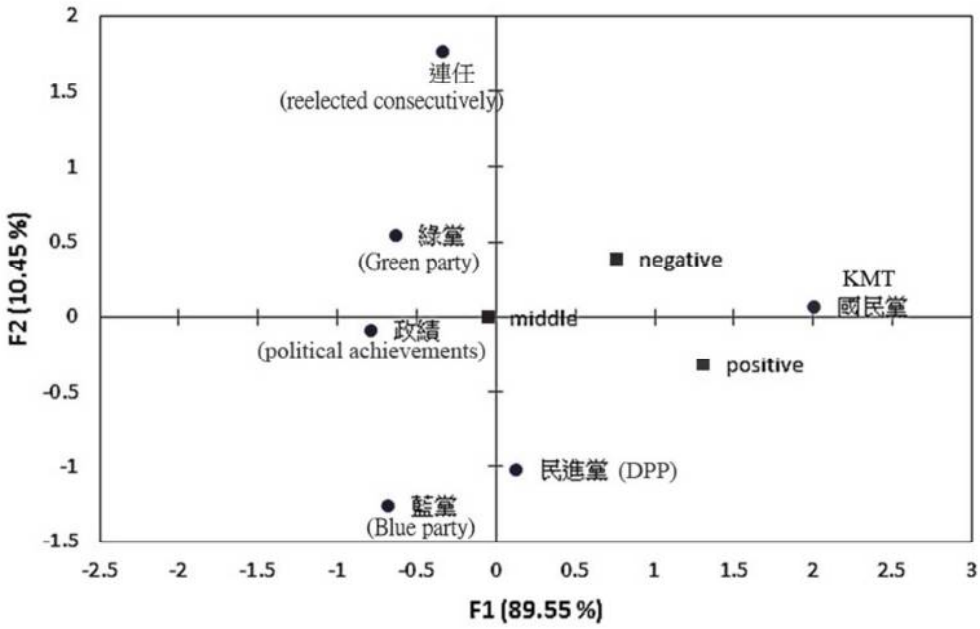

Figure 3 Correspondence analysis of Candidates' backgrounds for Hu.

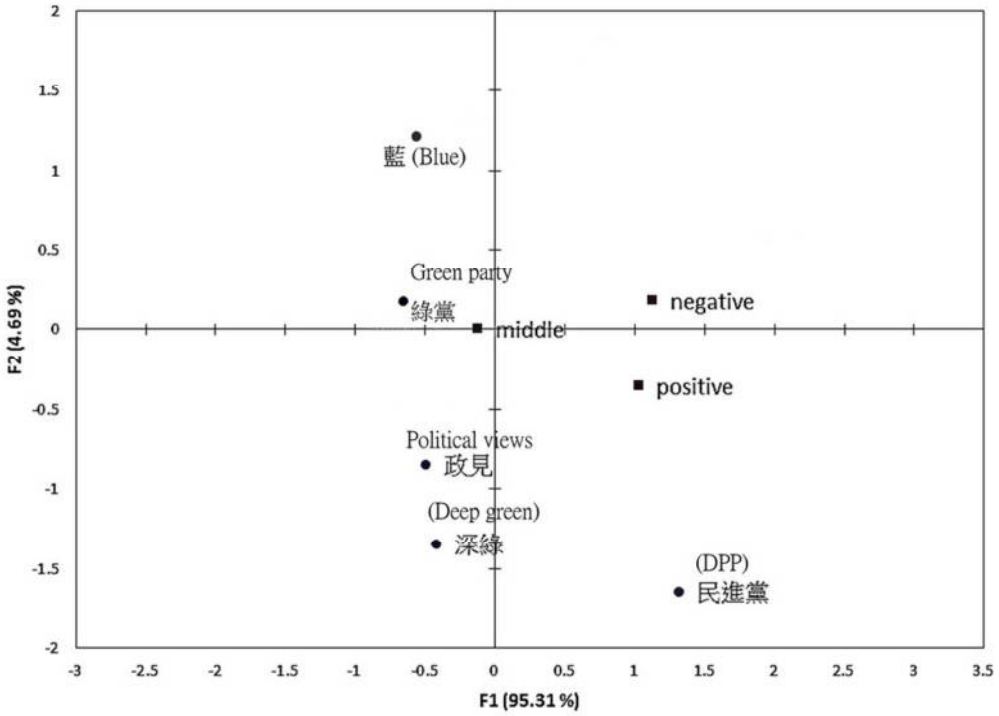

Figure 4 Correspondence analysis of Candidates' backgrounds for Lin.

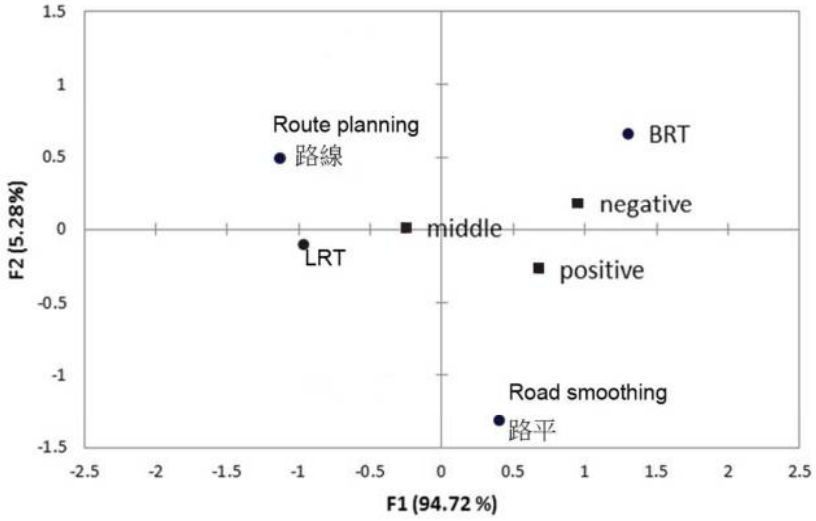

Figure 5 Correspondence analysis of transport infrastructure for Hu.

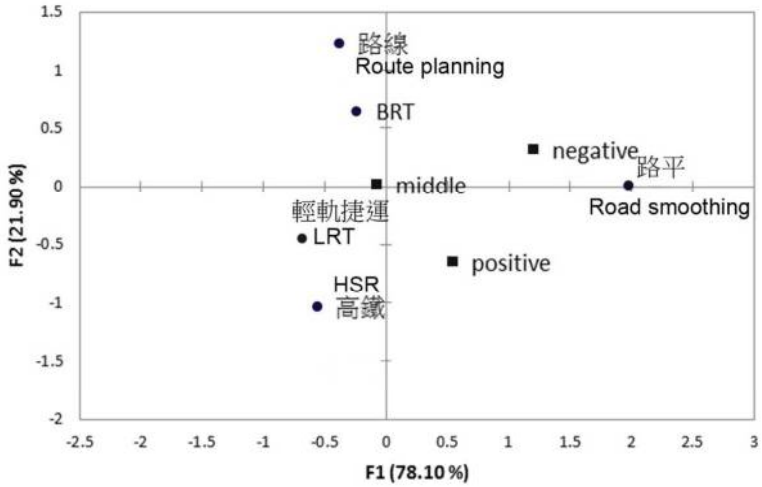

Figure 6 Correspondence analysis of transport infrastructure for Lin.

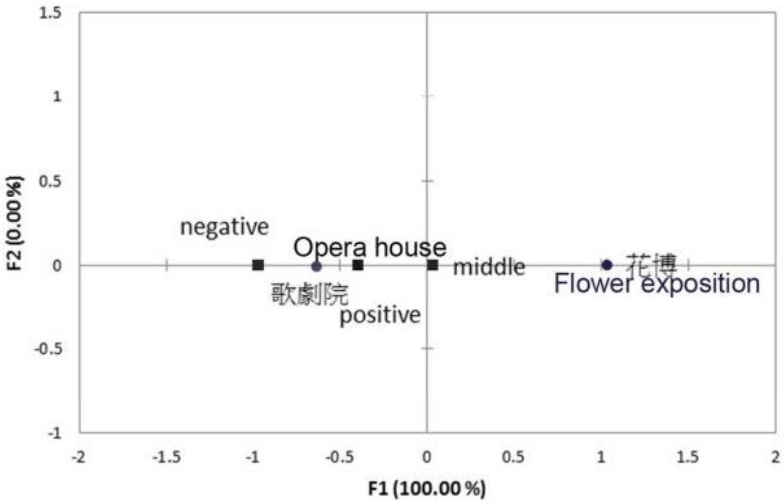

Figure 7 Correspondence analysis of arts and culture for Hu.

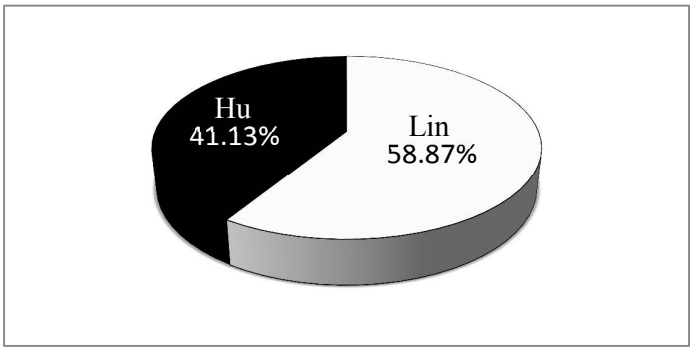

Figure 8 The proportions of the score of the two candidates.

Table 1 The age distribution of Internet access population in Taiwan in 2014.

Age	15~24	25~34	35~44	45~54	55~
Internet access (%)	24	29	24	14	9

Table 2 Opinion phrase extraction rules for identifying the opinion words.

item	Rules	Description
1	Na+(Vi/Vt+Vi/Vt)	The attribute word followed by (Vi/Vt+Vi), such as "政策 (policy, Na) "+"執行 (execute, Vi) "+"失敗 (fail, Vi)".
2	Na+(T+N)	The attribute word followed by (T+N) , such as "台灣 (Taiwan, Na) "+"的~'s; of; possessive particle, T"+"民主 (democracy, N) ".
3	Na+(ADV+Vi/Vt)	The attribute word followed by (ADV+Vi) , such as"治安 (public security, Na)"+"很 (very, ADV)"+"差 (bad, Vi)"
4	(ADV+Vi/Vt)+Na	(ADV+Vi/Vt) in front of the attribute word, such as "不 (didn't, ADV)" + "做 (do, Vt) "+"事 (anything, N)"
5	(N+ADV+Vi/Vt)+Na	(ADV+Vi/Vt) in front of an attribute word, such as "北京 (Beijing, N)+ "已然 (already, ADV)"+ "介入 (involve, Vt)" + "台灣 (Taiwan, Na) "
6	(Vi/Vt+T)+Na	(Vi/Vt+T) in front of an attribute word, such as "認真 (earnest, Vi)" +"的~'s; of; possessive particle, T"+"市長 (mayor, Na)".

Table 3 Negative words listed in the database.

無 (no)	不用 (need not)	枉 (futile)	徒然 (in vain)
不 (no/not)	何必 (why should)	未 (not /not yet)	瞎 (groundless)
否 (negate)	何須 (there is no need)	休 (don't)	虛 (empty)
勿 (don't)	無庸 (not need to)	別 (don't)	空 (hollow)
毋 (do not/no)	白白 (for nothing)	沒 (none)	非 (not)

Table 4 The numbers of positive comments of Hu (i=1) of topic CB (k=1) in each degree level

Degree level	Number of comments
1	22
2	54
3	14
4	44
5	28
6	70

Table 5 Summary of each topic by candidate.

	Topics	CB+	CB-	TI+	TI-	LP+	LP-	SW+	SW-	AC+	AC-	LC+	LC-
	k	1		2		3		4		5		6	
Hu	No. of Comments	232	184	118	136	14	4	16	4	10	12	2	0
	Percentage of degree words of extreme level	30	20	5	13	0	1	0	1	0	0	0	0
	Score	435	305.6	137.6	187.2	1.8	1	1	0.6	1	0.8	0	0
Lin	No. of Comments	334	288	38	52	0	4	0	4	0	0	0	0
	Percentage of degree words of extreme level	17	12	5	15	0	0	0	0	0	0	0	0
	Score	527.8	385.2	39.2	65.6	0	0.2	0	0.2	0	0	0	0