Journal of Informetrics 10 (2016) 634–644

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Research article

Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling

Hyui Geon Yoon[a], Hyungjun Kim[a], Chang Ouk Kim[a,*], Min Song[b]

[a] Department of Industrial Engineering, Yonsei University, Seoul, Republic of Korea
[b] Department of Library and Information Science, Yonsei University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 23 July 2015
Received in revised form 30 March 2016
Accepted 30 March 2016

Keywords:
Sentiment analysis
Shrinkage regression
Topic modeling

ABSTRACT

We propose a method to analyze public opinion about political issues online by automatically detecting polarity in Twitter data. Previous studies have focused on the polarity classification of individual tweets. However, to understand the direction of public opinion on a political issue, it is important to analyze the degree of polarity on the major topics at the center of the discussion in addition to the individual tweets. The first stage of the proposed method detects polarity in tweets using the Lasso and Ridge models of shrinkage regression. The models are beneficial in that the regression results provide sentiment scores for the terms that appear in tweets. The second stage identifies the major topics via a latent Dirichlet analysis (LDA) topic model and estimates the degree of polarity on the LDA topics using term sentiment scores. To the best of our knowledge, our study is the first to predict the polarities of public opinion on topics in this manner. We conducted an experiment on a mayoral election in Seoul, South Korea and compared the total detection accuracy of the regression models with five support vector machine (SVM) models with different numbers of input terms selected by a feature selection algorithm. The results indicated that the performance of the Ridge model was approximately 7% higher on average than that of the SVM models. Additionally, the degree of polarity on the LDA topics estimated using the proposed method was compared with actual public opinion responses. The results showed that the polarity detection accuracy of the Lasso model was 83%, indicating that the proposed method was valid in most cases.

1. Introduction

Social network service (SNS) provides a new platform on which users can freely express their opinions. Users share their opinions on a variety of issues through SNS platforms, and various researchers have conducted analyses focusing on the “exchange of opinions” that occurs on SNS platforms. Twitter is a representative SNS platform. Twitter’s most prominent feature is that users worldwide can communicate quickly in 140 or fewer characters (Jansen, Zhang, Sobel, & Chowdury, 2009), and features such as “following”, “followers”, and “retweets” enable the rapid exchange of information. Some have argued that Twitter played a role in the election of U.S. President Barack Obama (Cogburn & Espinoza-Vasquez, 2011). Furthermore, Larsson and Moe (2012) studied the effect of Twitter on Swedish election campaigns, and Graham, Broersma, Hazelhoff, and van’t Haar (2013), investigated how Twitter was used by candidates in the U.K. general election.

* Corresponding author.
E-mail address: kimco@yonsei.ac.kr (C.O. Kim).

http://dx.doi.org/10.1016/j.joi.2016.03.006
1751-1577/

Existing polarity (or sentiment) analysis focuses on classifying the polarity of individual texts (e.g., web reviews or tweets) by selecting important features through methods such as n-grams (Kennedy & Inkpen, 2006; Kouloumpis, Wilson, & Moore, 2011; Pak & Paroubek, 2010; Pang, Lee, & Vaithyanathan, 2002), word subsequence (Matsumoto, Takamura, & Okumura, 2005; Xia, Zong, & Li, 2011), information gain (Forman, 2003; Zhang, Ye, Zhang, & Li, 2011), and recursive feature elimination (Abbasi, Chen, & Salem, 2008; Abbasi, Chen, Thoms, & Fu, 2008). A tweet is then classified via algorithms, such as the naïve Bayes (Xia et al., 2011), maximum entropy (Pang et al., 2002; Xia et al., 2011), or support vector machine (SVM) algorithms (Abbasi, Chen, Salem, 2008; Dang, Zhang, & Chen, 2010; Moraes, Valiati, & Neto, 2013; Saleh, Martín-Valdivia, Montejo-Ráez, & Ureña-López, 2011). However, to understand the direction of public opinion on a political issue, it is important to analyze the degree of polarity on the major topics at the center of the discussion in addition to the individual tweets. For example, it is nearly impossible to identify all of the individual opinions of the public regarding a policy presented by a candidate during an election.

This study proposes a method to analyze public opinion regarding a specific political issue (or event) by utilizing polarity analysis of Twitter data at the individual tweet level as well as at the topic level. It is expected that when properly used, polarity analysis of social media can provide several benefits, including (1) reducing the cost of traditionally conducted public opinion polls, (2) augmenting such public opinion polls, and (3) producing a more accurate gauge of public opinion overall. In addition, analyzing whether a policy proposed by a candidate is viewed favorably can help candidates develop future policies.

The proposed method consists of two stages. The first stage detects polarity in tweets using shrinkage types of multivariate regression models, namely, Lasso and Ridge (Tibshirani, 1996). We define the status of an opinion as positive or negative, excluding a neutral opinion. Twitter data reflect opinions regarding candidates that were confirmed directly, and supervised learning was conducted based on these data. The dependent variable for the polarity response is set to one when support is expressed for a certain candidate and zero when an opposing opinion is introduced; furthermore, each term in tweets is set as an independent variable to examine the effect of each term on tweet polarity. The regression coefficients of the Lasso and Ridge models indicate the degree of the term's impact on the polarity response, and we consider the regression coefficients as the sentiment scores for the terms. Lasso and Ridge added a penalty section to perform a variable reduction that maximizes detection performance—the terms that are determined to be influential on the response have non-zero regression coefficients, whereas insignificant terms have coefficients near zero. Compared with SVMs (Cortes & Vapnik, 1995), which are frequently used for polarity analyses, the regression model is able to estimate the sentiment scores of the terms in the tweets.

Although the polarity of individual tweets can be detected, the results only reveal the number of opinionated tweets; they do not show what political topics people express their opinion on. Policy makers want to know the public responses to a political issue. Therefore, in addition to the polarity detection of individual tweets, it is important to capture the major topics on which public opinion is expressed and detect the polarity of such opinions. Thus, the second stage assigns a polarity code to the major topics under discussion rather than the individual tweets. In this step, a latent Dirichlet analysis (LDA) model (Blei, Ng, & Jordan, 2003) is applied to identify the major topics. LDA represents a topic as a group of terms that frequently occur in the texts under analysis. For each topic, polarity is estimated via the sentiment scores of the terms included in the group. To the best of our knowledge, our research is the first to estimate the polarities of public opinion derived from LDA topics in this manner.

An experiment using the proposed method was conducted on the 2014 mayoral election in Seoul, South Korea, which occurred on June 4, 2014. We compared the total polarity detection accuracy of the regression models with five SVM models with different numbers of input terms selected by a feature selection algorithm. The results indicated that the performance of the Ridge model was better than that of the SVM models. Additionally, the polarity of LDA topics estimated using the proposed method was compared with actual public opinion responses. The results showed that the polarity detection of the Lasso method was valid in most cases.

The remainder of this paper is organized as follows. Section 2 reviews the prior studies related to opinion polarity analysis; Section 3 details the proposed method, which includes a data-preprocessing step; Section 4 introduces the Twitter data related to a mayoral election in Korea and presents the analysis results; and Section 5 presents the conclusion of this study and discusses directions for future research.

2. Related works

Communicated ideas and intentions are captured by language, and we use various expressions to reveal such ideas and intentions. The basic units that constitute this expression are terms, or morphemes, and it is through a collection of these concepts that various sentences are formed and further developed to express opinions (Na, Lee, Nam, & Lee, 2009; Zhang & Liu, 2011). Even in the same context, the choice of terms expresses differences in opinions regarding a specific issue or event. Thus, it is reasonable to hypothesize that the polarity of a term can determine the polarity of the speaker's opinion regarding a specific issue. It is presumed that when negative terms (i.e., terms with negative polarizing characteristics) are used to describe an opinion regarding a specific issue, this usage can polarize the speaker's opinion of this issue in a negative direction (Breck, Choi, & Cardie, 2007).

Early research related to the field of opinion dynamics proposed models of simple opinion evolution processes, beginning with the Ising model, which is based on statistical thermodynamics and quantum mechanics spin concepts (Galam, Gefen,

& Shapir, 1982; Galam & Moscovici, 1991). The basic concept of the Ising model limits an individual's opinion regarding a specific object to two states. Although a complex human may feel resistant to the idea that there can only be two opinion states for each object, a decision maker typically has only a few choices in a decision-related action, such as an election. Wegrzyn-Wolska and Bougueroua (2012) used both social network analysis techniques and a text mining method to investigate opinion polarity in the 2012 French presidential election. Similarly, Nooralahzadeh, Arunachalam, and Chiru (2013) have also studied time-series polarity in U.S. and French elections.

Traditional polarity analysis uses simple feature selection to classify the polarity of individual texts. Pang et al. (2002) extracted unigrams, bigrams, and adjectives from movie reviews as features and applied three machine learning algorithms (naïve Bayes, maximum entropy, and SVM) to classify the polarity of an individual movie review. Kennedy and Inkpen (2006) used the appearance frequency of n-grams and adjectives in texts using the term-counting method. Pak and Paroubek (2010) collected a Twitter sample for polarity analysis and constructed a set of n-grams (unigrams, bigrams, and trigrams) to investigate the effects of consecutive words. Kouloumpis et al. (2011) applied n-gram, lexicon (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011; Wilson, Wiebe, & Hoffmann, 2009), part-of-speech, and emoticons (Hogenboom et al., 2013).

Matsumoto et al. (2005) proposed using word subsequence and dependency trees to express the correlations between words. They used the correlations as features for an SVM. Xia et al. (2011) introduced an ensemble of feature sets to represent various features. Forman (2003) and Zhang et al. (2011) used information gain to select effective features. Information gain is an algorithm that measures the difference in the classification results when individual features are used for the text classification. One beneficial characteristic of information gain is that it can consider class labels (e.g., polarity codes) when selecting features. This characteristic can significantly improve classification performance. Abbasi, Chen, Salem (2008) and Abbasi, Chen, Thoms et al. (2008) used recursive feature elimination for feature selection (Guyon, Weston, Barnhill, & Vapnik, 2002). The SVM is a powerful classification algorithm; many researchers have employed SVMs as a polarity classifier (Abbasi, Chen, Salem, 2008; Dang et al., 2010; Moraes et al., 2013; Saleh et al., 2011). In addition, researchers have employed other classification models such as ensemble (Zhang, Ma, Yi, Niu, & Xu, 2015), decision tree (Prasad, Kumar, Prabhakar, & Pal, 2015), and the logistic regression-based binary classifier (Bollegala, Weir, & Carroll, 2013).

Feature selection and classification have been conducted separately in related studies. If the feature selection method is not effective for a certain classification algorithm, the classification may exhibit poor performance. In this study, we employ shrinkage regression, which performs feature selection and classification at the same time. In addition, it is able to estimate the sentiment scores of the terms in the tweets compared with other feature selection or classification algorithms.

Detecting the polarity of individual tweets is important, but the simple aggregation of the polarity of individual tweets does not necessarily reveal public opinion on important issues or agendas. If we only detect the polarity of individual tweets, it may be possible to deduce the overall polarity. However, we cannot analyze which topics are positive or negative. Without understanding topic-specific public opinion, policy makers who make proposals do not know which policies are considered to be helpful. In other words, it is difficult to propose policies that properly reflect public opinion. Thus, capturing major topics on Twitter and detecting the degree of polarity on such topics can provide important information for policy makers. This study proposes a method that detects the polarity of public opinion.

3. Proposed method

3.1. Preprocessing

A technique that splits sentences into terms or morphemes has been studied in the fields of language and linguistic study; this technique is called tokenization. Part-of-speech (POS)-tagging classifies tokenized terms as nouns, verbs, etc. (Manning, Raghavan, & Schütze, 2008) and has been applied to Twitter data (Gerber, 2014; Laboreiro, Sarmento, Teixeira, & Oliveira, 2010). It uses an algorithm that decomposes text data into nouns, postpositions, and verbs. Text mining packages, such as Mallet (McCallum, 2002), contain POS-tagging functions for texts in English but do not offer the POS-tagging function for texts in other languages, such as Korean tweets, which were used in the experiment. Therefore, this study uses a different general-purpose package, namely, the "KOMORAN morphological analyzer" (KOMORAN, 2003), to conduct POS-tagging on Korean text. Other POS types are removed, and only nouns are used in this study.

A training dataset is needed to build a regression model. Table 1 shows an example training dataset. In the table, the independent variables are terms of all tweets when the number of all terms in the training dataset is N. The independent variables have a value of one or zero: one was assigned to the terms included in each tweet, and zero was assigned to those that were not included in a tweet. Polarity codes should be manually assigned to the training tweets to complete the training dataset. The dependent variable is binary-valued such that tweets with an opinion that supports a specific candidate have a polarity code of one, and tweets opposing a specific candidate are assigned a code of zero. If a tweet does not contain a particular opinion regarding a candidate, the tweet cannot be used in the analysis and is disregarded. Three of the authors read the tweets and assigned polarity codes to the tweets in the training datasets. For each tweet, the authors declared the polarity codes individually and voted for the final decision.

Table 1
Preprocessing of Twitter data to apply the shrinkage regression.

Tweet No.	Independent variable						Dependent variable
	Term 1	Term 2	Term 3	Term 4	...	Term *N*	Polarity code
1	0	1	0	0	...	0	0
2	0	0	0	1	...	0	0
3	1	0	0	1	...	0	0
4	1	0	0	0	...	1	1
5	0	0	0	1	...	0	0
6	0	0	1	0	...	0	0
7	0	1	0	0	...	1	1
8	1	0	0	0	...	0	1

3.2. Polarity detection of tweets

Regression analysis using the Lasso and Ridge models is conducted to classify tweet polarity. Both models add a penalty section to determine a strategy for expanding the influence of the important independent variables for polarity classification. The difference between the two models is that each uses a different variable reduction method. The important variables are retained in the Lasso model, while the rest are excluded. In contrast, in the Ridge model, although the coefficients of the non-important variables can be reduced to a value close to zero, the corresponding variables cannot be complfetely removed by making them zero. Variable reduction levels are determined according to the penalty parameters included in each model. Regression coefficients that satisfy the objective functions in Eqs. (1) and (2) are estimated for the Lasso and Ridge models, respectively. Here, β_j is the regression coefficient; λ signifies the parameters of the penalty section; and p indicates the number of independent variables.

$$\text{Minimize}\, l\left(\beta_0, \beta_1, .., \beta_p\right) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{1}$$

$$\text{Minimize}\, l\left(\beta_0, \beta_1, .., \beta_p\right) + \lambda \sum_{j=1}^{p} {\beta_j}^2. \tag{2}$$

In the above equations, $l\left(\beta_0, \beta_1, .., \beta_p\right)$ is the loss function, also called binomial deviance. It has a range of $[0, \infty)$. For the binary-valued response, the loss function is defined as in Eqs. (3) and (4) (Meier, Van De Geer, & Bühlmann, 2008). Here, x_{ij} is the independent variable and denotes term j of tweet i; x_i is the collection of terms in tweet i; y_i is the dependent variable and denotes the polarity of tweet i; and n indicates the amount of data.

$$l\left(\beta_0, \beta_1, .., \beta_p\right) = -\sum_{i=1}^{n} [y_i \times \log \mathrm{p}(x_i) + (1 - y_i) \times \log(1 - p(x_i))], \tag{3}$$

$$p(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}. \tag{4}$$

In Eq. (4), $p(x_i)$ implies the probability that a tweet is positive. When y_i is one (i.e., the tweet is positive), the loss function in Eq. (3) is reduced to $-\log p(x_i)$ and is minimized at $p(x_i) = 1$. When y_i is zero (i.e., the tweet is negative), the loss function in Eq. (3) becomes $-\log(1 - p(x_i))$, which is minimized at $p(x_i) = 0$. Thus, the Lasso and Ridge models are trained such that positive (or negative) tweets are classified correctly, minimizing the loss function in Eq. (3). For a new tweet that is not used in the model training, the trained models accept the POS-tagged tweet (only nouns are extracted) as the input and detect the polarity of the tweet by means of the output value of the dependent variable.

The disadvantage of a general logistic regression is that it can lead to multicollinearity when there are a large number of independent variables. In contrast, variable reduction methods, such as Lasso and Ridge, account for the correlation between variables during the process, thereby reducing the influence of non-important variables. Because of these characteristics, these two variable reduction methods present a natural tendency to resolve the multicollinearity issue. The use of the Lasso and Ridge models is appropriate for this study because a large number of independent variables are presented in the training tweets as a direct outcome of the inclusion of every term in the tweets as an independent variable.

3.3. Polarity detection of topics

The regression coefficient betas of the trained Lasso and Ridge models can be considered the sentiment scores for the terms, and the positive/negative sign indicates whether the terms are positive or negative regarding the political issue of interest. In addition to detecting the polarity of a tweet, those coefficients can be further utilized to analyze topic polarity.

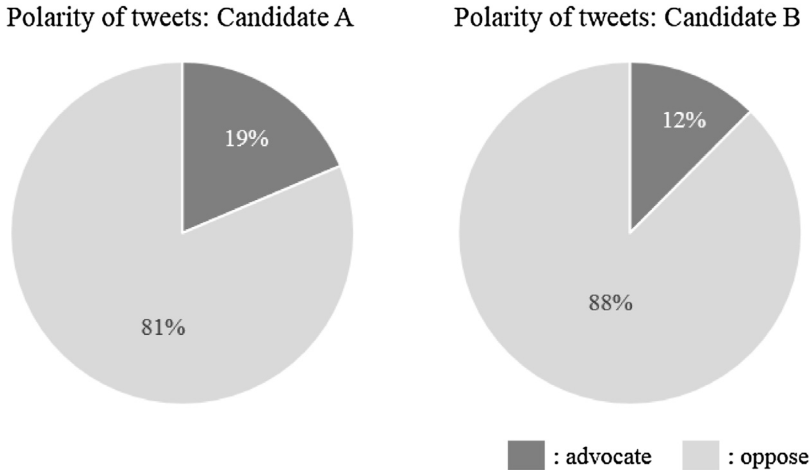

Fig. 1. Polarity ratio in tweets directed toward Candidates A and B.

For a political election, it is not sufficient for the candidates to predict the polarity of individual tweets. Instead, they must know the trends of the main topics or subjects that are actively discussed in public media. This study employs LDA to identify the main topics on Twitter.

LDA is a probabilistic topic model that assumes that documents (e.g., tweets) are characterized by mixtures of topics. In the context of LDA, a topic is composed of terms with creation probabilities. For each term position in a document, LDA identifies a topic, and the topic is composed of the terms included in the topic, measured probabilistically. Given a set of documents, LDA provides an algorithm that learns the topics and the terms associated with each topic. LDA requires one input parameter: the number of topics to extract. In this study, as shown in Eq. (5), we assign a topic's polarity code based on the sign of the sum of the sentiment scores for the terms that constitute the topic. The code F (or U) implies a favorable (or unfavorable) sentiment score. The probability of a term's occurrence could be added to Eq. (5). However, when we decide the names of the topics derived from LDA, particular terms are more significant, even if the probability of their term occurrence is low. In addition, if we add probability to the equation, the probability of some terms is too large, and the sentiment scores of other terms cannot be considered. For these reasons, the probability of term occurrence is not considered in Eq. (5).

$$\text{Polarity code of topic} = \text{sign}\left(\sum_{\text{term} \in \text{topic}} \text{sentiment score}_{\text{term}}\right), \tag{5}$$

$$\text{where sign}(x) = \begin{cases} \text{F} & \text{if} x \geq 0 \\ \text{U} & \text{otherwise} \end{cases}.$$

4. Experiment

4.1. Data

Twitter data related to the Seoul mayoral election were collected from approximately 1.5 million tweets posted from March 9 to June 3, 2014. The election was structured as a race between Candidate A from an opposition party and Candidate B from the ruling party. We identified that there were active discussions on Twitter from April 15 to April 30, 2014. The terms "Candidate A," "Candidate B," and "Seoul mayoral election" were set as keywords, and 11,703 tweets containing at least one of the three keywords generated during this period were extracted. From the sample, 1575 tweets were randomly selected and used as a training dataset to construct the regression models. To verify the performance of the proposed method, 408 tweets collected after April 30 were used to construct a test dataset. In total, 1575 tweets before May were included in the training dataset, and 408 tweets after May were included in the test dataset.

The polarity coding results indicated that of the 1575 tweets, there were only 269 and 308 tweets related to Candidates A and B, respectively. Many of the tweets did not show a distinct polarity and were thus removed from the analysis. As mentioned in Section 3.1, three of the authors read the tweets and assigned polarity codes to the tweets. For Candidate A, the coders agreed unanimously on the polarity codes of 96% of the 269 tweets and voted for final decisions regarding the remaining 4%. For Candidate B, they assigned an identical code to each of 98% of the 308 tweets and assigned codes to the remaining 2% throughout voting. As shown in Fig. 1, the majority of the responses directed toward both candidates on

Table 2
Summary of the Lasso and Ridge models.

Model	Candidate	Shrinkage regression	Tweets	Number of independent variables in the training dataset	Lambda	Loss function value	Selected number of independent variables
I	A	Lasso	269	1468	0.01	0.41	194
II	A	Ridge	269	1468	2.22	0.81	1468
III	B	Lasso	306	1307	0.01	0.19	168
IV	B	Ridge	306	1307	2.04	0.81	1307

Twitter were negative. Only 19% of the tweets related to Candidate A and 12% of the tweets related to Candidate B were supportive.

4.2. Polarity detection results of tweets

Lasso and Ridge models were generated for each candidate, yielding a total of four different models, as shown in Table 2 (i.e., Models I–IV). The small positive loss function values in Table 2 indicate a high goodness of fit for the model. The lambda values signify the parameters of the penalty section in the objective function in Eq. (1). A higher lambda value indicates that fewer terms are chosen as the independent variables in the regression analysis. The lambda value that minimizes the loss function was chosen in this experiment, and the best lambda values found for the four models are shown in Table 2. In the Lasso models I and III, 194 and 168 terms were selected for a lambda of 0.01, with minimum loss function values 0.41 and 0.19, respectively. In contrast, variable reduction was not observed in the Ridge models II and IV.

The detection accuracy of the 408 test tweets was measured to evaluate the performance of the four models. In addition, a SVM was introduced as a comparison model. The SVM is a well-known machine-learning algorithm that exhibits excellent performance in text classification problems. Three accuracy measures, shown in Eqs. (6)–(8), were used to compare the performance of the models. In particular, Eqs. (7) and (8) measure the accuracy of positive and negative tweets, respectively. The positive and negative tweet accuracies were measured because as noted in Section 4.1, negative tweets constituted approximately 85% of the training tweets overall, indicating that a tweet is five times more likely to be negative than positive. In this case, a prediction model can exhibit an accuracy of 85% even if the model declares all tweets to be negative. Therefore, the accuracy of positive and negative tweets should be evaluated separately in addition to the total accuracy.

$$\text{Total accuracy} = \frac{\#(\text{positive tweets predicted positive}) + \#(\text{negative tweets predicted negative})}{\#(\text{all tweets})}, \tag{6}$$

$$\text{Positive tweet accuracy} = \frac{\#(\text{positive tweets predicted positive})}{\#(\text{positive tweets})}, \tag{7}$$

$$\text{Negative tweet accuracy} = \frac{\#(\text{negative tweets predicted negative})}{\#(\text{negative tweets})}. \tag{8}$$

For the 408 test tweets, Table 3 provides a comparison between the results of the polarity coding by the authors and the prediction results based on the Lasso, Ridge, and SVM models. In terms of total accuracy, the detection accuracy tended to be higher in the Ridge models compared with the Lasso models. When comparing the results of models I and II with regard to the 269 tweets to which the polarity code related to Candidate A was assigned, model II exhibited 3% higher accuracy. The predictions made by models III and IV with regard to the 308 tweets related to Candidate B also indicate that the Ridge model produced better results, with a difference of 2%. From these results, we can confirm that the Ridge model is a better polarity detection model than the Lasso model in the experiment. However, in most cases, the accuracy is lower than 85% because an oversampling technique was used in building the regression models and the SVMs to set the proportion of positive and negative tweets at 50% each. In other words, positive tweets were oversampled to generate similar numbers of positive and negative tweets.

The SVM results with feature selection are also shown in Table 3. We used a recursive feature elimination algorithm for the feature selection (Guyon et al., 2002). The Ridge and Lasso models include feature selection, so feature selection is also needed for the SVM for a fair comparison. The algorithm requires the number of terms to be selected as the input. In this

Table 3
Accuracy of the Lasso, Ridge, and SVM models.

Prediction accuracy (%)	Lasso		**Ridge**		SVM-100		SVM-200		SVM-500		SVM-1000		SVM-Full	
	A	B	**A**	**B**	A	B	A	B	A	B	A	B	A	B
Total accuracy	75	81	**78**	**83**	75	81	74	84	70	82	66	72	62	68
Positive tweet accuracy	44	46	**36**	**49**	41	37	48	57	67	49	73	51	76	46
Negative tweet accuracy	85	85	**90**	**88**	85	87	67	87	70	87	64	75	58	71

Bold values represent the best performance scores.

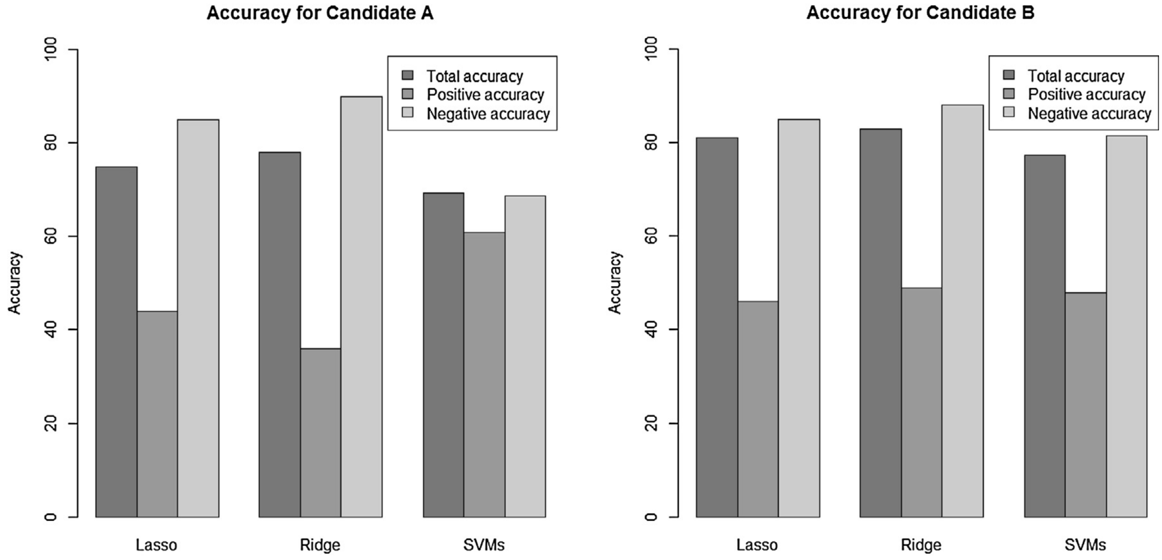

Fig. 2. Bar charts of the Lasso, Ridge, and SVM models.

study, the number of terms used to train the SVM was set at 100, 200, 500, and 1000. In addition, an SVM model using all of the terms (1468 terms for Candidate A and 1307 terms for Candidate B) in the training dataset was also tested. The results indicate that the performance of the Ridge model was approximately 7% better on average than that of the five SVM models. The performance of the SVM models depended on the number of terms used as inputs; the SVM trained with all of the terms in the trained dataset showed the worst performance in terms of total accuracy, whereas the SVM with 200 terms resulted in the best performance. The Ridge model was approximately 2.5% more accurate on average than the SVM with 200 terms. In terms of negative tweet accuracy, the Ridge model tended to detect negative tweets more accurately. In contrast, the SVM model tended to detect positive tweets more accurately.

A summary of the polarity detection results of the tweets is shown in Fig. 2. The *x*-axis is the Lasso, Ridge, and SVMs; it denotes the average accuracy of all of the SVM models. With respect to total accuracy and negative tweet accuracy, the Ridge model performed best. The regression models are appropriate for polarity detection because these models are capable of not only performing polarity detection similarly to the SVM but also of determining the sentiment scores for the terms. The SVM does not provide sentiment scores for the terms in the tweets.

4.3. *Sentiment scores for the terms in tweets*

Table 4 lists the 15 key supportive and opposing terms derived from the terms' regression coefficient betas. The Korean terms were translated into English and are shown in the table. The intercepts (β_0) of the Lasso regression models (models I and III), marked in dark gray on the table, have negative values. These results are related to the aforementioned overall trend of the tweets. An intercept with a negative value indicates that the public opinion is negative. Thus, the tweets are more likely to contain an opposing opinion than a supportive opinion.

Terms in light gray are those that the authors considered important for determining the polarity of a tweet. For example, "Mayor", which is the fifth supportive term in model I, occurs in numerous tweets mentioning that Candidate A should be re-elected as mayor of Seoul. Therefore, the term "Mayor" can be considered to be a supportive term for Candidate A. With regard to the seventh supportive term in model I, "Ability", there were many tweets stating that the ability of Candidate A is outstanding because he reduced Seoul's large debt.

In Table 4, the Lasso models (i.e., models I and III) present the top 15 supportive and opposing terms (highlighted in light gray) more distinctly compared with the Ridge models (i.e., models II and IV). This difference results from the primary model characteristics of the Lasso and Ridge models. Because the Ridge model cannot set the regression coefficients to zero unless the penalty parameter lambda is set to infinity, the model cannot produce clear differences among the coefficients. Consequently, the Lasso model shows slightly better performance in detecting the polarity of terms than the Ridge model with regard to the total accuracy in Table 3 and the number of highlighted terms in Table 4.

There are two advantages of this method. First, it is relatively simple. A shrinkage regression is used to calculate the sentiment scores for terms in tweets while detecting the polarity of the tweets. This method could supplement the shortcomings of other methods, which require human knowledge or involve a complicated process (Tsai, Wu, Tsai, & Hsu, 2013). Second, the sentiment scores are adjusted based on the analysis domain: some terms have different sentiment scores in the different domains. For example, "cancer" is usually a negative term, but it was considered to be positive term in models I and II. In reality, the term "cancer" was used in the sense of "cured of cancer" in many of the tweets.

Table 4
Fifteen major supportive and opposing terms derived by each model.

Model I		Model II		Model III		Model IV	
Supportive Terms							
Term	Coefficient	Term	Coefficient	Term	Coefficient	Term	Coefficient
Cancer	5.23	Cancer	0.20	The older generation	7.36	Fighting	0.22
M. B. Lee	5.21	Purity	0.19	Fighting	6.61	The truth	0.19
Fuse	4.84	Excellence	0.19	The truth	6.21	Atmosphere	0.19
Commandment	4.47	Leader	0.18	Curve	6.20	Complement	0.19
Mayor	3.99	Excuse	0.18	Support	6.20	Distortion	0.19
Solace	3.52	The bereaved	0.18	Complement	6.11	Curve	0.18
Ability	3.07	Love	0.18	Reason	5.93	Increase	0.18
Love	1.98	Competence	0.15	The press	5.76	Story	0.17
Today	1.97	M. B. Lee	0.15	Transportation	5.72	Generosity	0.17
Emperor	1.83	Vote	0.15	Composition	5.61	The older generation	0.17
Admiration	1.77	Fund	0.15	Officer	5.22	Witch hunt	0.17
Government	1.75	Hiking	0.14	Deploration	5.11	Hyena	0.17
J. I. Moon	1.61	Native	0.14	Speech	5.09	Relief	0.16
Leader	1.61	Disgust	0.14	Relief	4.35	Thief	0.16
Jindo Island	1.59	Logic	0.14	Crisis	3.92	Officer	0.15
Opposing Terms							
Term	Coefficient	Term	Coefficient	Term	Coefficient	Term	Coefficient
Blue House	-6.97	Blue House	-0.27	Current	-5.23	Saint	-0.23
Seongnam	-4.55	The bereaved	-0.20	Treatment	-4.68	Treatment	-0.23
Son	-4.08	Difference	-0.20	Celebrity	-4.39	Possibility	-0.21
New daily (online newspaper)	-4.04	Generous	-0.20	Event	-4.12	Defamation	-0.21
Start	-3.32	Discontinuance	-0.19	Predetermination	-3.64	Memory	-0.21
Party	-3.21	Seoul Floating Island	-0.19	J. Y. Jeong	-3.57	Bless	-0.20
H. H. Lee	-3.08	Reappointment	-0.19	Luck	-2.64	J. Y. Jeong	-0.20
Iron will	-3.01	City hall	-0.18	(Intercept)	-2.64	Astonishment	-0.20
Game room	-2.90	Joke	-0.18	Bless	-2.23	Power	-0.20
Concern	-2.69	Defection	-0.18	Uncivilized	-2.06	Carbon copy	-0.20
H. S. Kim	-2.53	Shadow	-0.18	Today	-1.85	Nervousness	-0.19
(Intercept)	-2.43	Dark clouds	-0.18	Resignation	-1.85	Dumbfounded	-0.19
Accident	-1.76	Obesity	-0.18	Word	-1.71	Dissonance	-0.19
Defamation	-1.64	Shape	-0.18	Son	-1.56	Son	-0.19
Pro-North Koreans	-1.17	Excuse	-0.17	H Heavy Industries	-1.49	City hall	-0.19

4.4. Polarity detection results of topics

In this paper, LDA was used to identify the main topics in the training dataset. In the experiment, we attempted to extract various topic numbers (from five to fifty topics) from the Twitter data, and we found that ten topics were suitable for the election situation. If we used more than ten topics, there were too many insignificant and overlapping topics. Gibbs sampling with 2000 iterations was used to estimate the distributions to be used in the LDA. Additionally, 30 top terms were established to describe each topic. We provided topic names after examining the terms that constituted the topics. Table 5 lists the LDA topics and the polarity codes of the topics predicted using Eq. (5). The ten top terms in the LDA topics are shown in Appendix A. In addition, the authors assigned a polarity code to each topic based on public opinion—code "F" (Favorable) means that the topic is strongly supportive of the candidate; code "U" (Unfavorable) means that the topic is distinctly opposing of the candidate. The authors' coding results are shown in the last two columns of Table 5. Similar to the coding of the individual tweets, three of the authors assigned polarity codes to the topics and voted for the final decision.

The shaded codes indicate the cases in which the results predicted using the Lasso and Ridge models were different from the results determined by the judgment of the authors. Three cases using the Lasso model and five cases using the Ridge model were different from the judgment of the authors. The detection accuracies for the Lasso and Ridge models were 83%

Table 5
Polarity of topics using Lasso and Ridge models.

No.	Topic	Lasso		Ridge		Judgment of authors	
	Candidate	A	B	A	B	A	B
1	Criticism of Candidate B	F	F	F	U	F	U
2	Criticism of Candidate A	U	U	U	U	U	F
3	Sewol ferry disaster	U	F	U	F	U	F
4	Candidate B winning in the polls	U	F	U	U	U	F
5	Insignificant topic	U	F	U	F	-	-
6	Criticism of Candidate B's son regarding the Sewol ferry disaster	F	U	U	U	F	U
7	Candidates from various regions	U	F	U	F	U	F
8	Encouragement for Candidate A	F	F	U	U	F	U
9	Military service evasion of Candidate A's son (1)	U	F	U	U	U	F
10	Military service evasion of Candidate A's son (2)	U	F	U	U	U	F

(15/18) and 72% (13/18), respectively. A more detailed analysis of the results indicates that the prediction for Candidate B on Topic 1 was incorrect for the Lasso model; however, the prediction for Candidate B on Topic 2 was not entirely wrong. Topic 2 is criticism of Candidate A, and this criticism was not helpful for Candidate B in the actual election. In fact, the sum of the sentiment scores for the terms that constitute Topic 2 for Candidate B was a small negative value of −0.79. This result implies that using the sentiment scores for learned terms in topic polarity analysis can complement human analysis by identifying parts that humans can miss.

Topic 1 is the only topic wherein the Ridge model shows a superior polarity estimation compared with the Lasso model. Topic 1, "Criticism of Candidate B," can be considered to be unfavorable for Candidate B; however, the Lasso model indicates a positive sum of the sentiment scores for Candidate B. Excluding this case, the Lasso model exhibited a better result than the Ridge model. Therefore, we recommend the use of the Lasso model when predicting the polarity of topics. Because of the nature of the models, whereas large variations in the sentiment scores for the terms were not observed in the Ridge model, the Lasso model presents clear differences and thus provides better results. Furthermore, the predicted polarities could be used by both candidates. For example, Topics 9 and 10 are both criticisms of Candidate A. The repetition of an identical topic signifies that there was a significant amount of public opinion on Twitter criticizing Candidate A. In light of such sentiments, Candidate A could have quickly identified new topics to try to shift public opinion, whereas Candidate B needed to use such topics to create an advantage during the campaign.

Polarity detection can be used in many areas where public responses are critical. In the area of political campaigns, candidates want to share what people think about their political proposals or behavior. Thus, the proposed polarity detection approach allows for the examination of public opinion polls in an inexpensive and timely manner. Candidates may want to determine the degree of polarity of the public and, by extension, which political proposals are considered favorably. The proposed approach that comprises two-tier polarity detection not only helps to identify individual opinions but also enables a specific understanding of the public's opinion about a particular theme in a political campaign.

5. Conclusion

This study proposed a method for detecting the polarity of tweets by employing two regression models: the Lasso and Ridge models. This method was applied to an analysis of tweets concerning the 2014 mayoral election in Seoul, South Korea. In the regression analysis, the terms included in the tweets were regarded as independent variables, and tweet polarity was treated as a dependent variable. Training tweets sampled in the active discussion period were used to build the regression models. The regression models can be used to predict the polarity in tweets that are not directly coded. We computed the prediction accuracy of the models using test tweets gathered after the active discussion period, and the results indicated that the prediction accuracy of the Ridge model was approximately 7% higher on average than that of five SVM models with feature selection. The regression models were capable of learning the sentiment scores of the terms contained in the training tweets, whereas the SVM could not. Topic polarity computed based on the terms' sentiment scores was also measured. Major topics related to the mayoral election were established by LDA, and the polarity scores for the topics were estimated.

We expect that the method proposed in this paper can be applied to identify public opinion about political campaigns as well as to various other domains that require the identification of public opinion, such as policy proposals or consumer responses to company products. Polarity detection of individual tweets helps to understand the aggregated opinion of the public at a high level. Conversely, the polarity detection of topics can help to identify topic-specific public opinion. The proposed approach allows for the detection of polarity in a systematic fashion. As demonstrated in the experiment, the proposed approach supplemented human mistakes. In addition, it could help candidates in political campaigns prepare their campaigns for future directions or strategies.

The disadvantage inherent in this method of determining topic polarity is that it does not consider the relationship between terms. A tweet consists of terms that generate meaning through the interactions between each term rather than being mere lists of terms. A more detailed understanding regarding topic polarity can be expected when methods such as the Bayesian network (Pelikan, 2005) and ontology (Euzenat and Shvaiko, 2007) are applied to indicate that such relationships exist among terms.

Author contributions

Hyui Geon Yoon: conceived and designed the analysis, collected the data, contributed data or analysis tools, performed the analysis and performed the regression analysis.

Hyungjun Kim: concieved and designed the analysis, collected the data, contributed data analysis tools, performed the analysis and performed the regression analysis.

Chang Ouk Kim: concieved and designed the analysis, collected the data, contributed data analysis tools, performed the analysis and assisted in writing the paper.

Min Song: concieved and designed the analysis, collected the data, contributed data analysis tools, performed the analysis and discussed the idea of this paper with the third author (corresponding author) and assisted in writing the literature review significantly.

Acknowledgments

This work was supported by the Yonsei University Future-leading Research Initiative of 2014 (2014-22-0116). In addition, this work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A3A2046711).

Appendix A. Major terms of each topic in Table 5

The major terms used within each topic in Table 5 are shown below. These terms were chosen based on their importance and the number of appearances of the terms.

No.	Topic	Ten major terms				
1	Criticism of Candidate B	Candidate B	Heavy industry	Seoul mayor	Worker	H.S. Kim
		Candidate A	Congressman	Subway	Bus fee	Temporary worker
2	Criticism of Candidate A	Candidate A	Seoul mayor	Military scandal	Election law	Candidate A's son
		Lawyer	Monkey	Lie	Democracy	Public official
3	Sewol ferry disaster	Candidate B	Son	Sewol	Ferry	Family
		Missing person	President	Accident	Disaster	Person
4	Candidate B winning in the polls	Candidate B	Seoul mayor	H.S. Kim	Candidate A	H.H. Lee
		Saenuri party	Opinion poll	G.H. Park	C.S. Ahn	Local election
5	Insignificant topic	Candidate A	Candidate B	Saenuri	Seoul mayor	President
		Superintend-ent of education	Incheon mayor	Umyeonsan (Mt.)	Landslide	Contribution
6	Criticism of Candidate B's son regarding the Sewol ferry disaster	Candidate B	Candidate A	Missing person	Uncivilized	Father
		Korea	Facebook	The bereaved	Saenuri	Son
7	Candidates from various regions	Candidate A	Seoul mayor	J.M. Lee	S.G. Kim	Superintendent of education
		Y.G. Song	H.J. Ahn	H.H. Lee	M.S. Choi	Local election
8	Encouragement for Candidate A	Candidate A	J.I. Moon	C.S. Ahn	Seoul mayor	Politician
		Judicial Research and Training Institute	Fuse	Lawyer	Supporter	Report
9	Military service evasion of Candidate A's son (1)	Candidate A	Seoul mayor	Candidate B	Military scandal	M.B. Lee
		Seoul Floating Island	Office	Korea	Subway	Son
10	Military service evasion of Candidate A's son (2)	Candidate A	Candidate A's son	Military scandal	Physical examination	Military manpower administration
		Press conference	Dr. Oh	Certification of contents	Deputy	Defamation

References

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems (TOIS)*, *26*(3), 12–34.

Abbasi, A., Chen, H., Thoms, S., & Fu, T. (2008). Affect analysis of web forums and blogs using correlation ensembles. *IEEE Transactions on Knowledge and Data Engineering*, *20*(9), 1168–1180.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bollegala, D., Weir, D., & Carroll, J. (2013). Cross-domain sentiment classification using a sentiment sensitive thesaurus. *IEEE Transactions on Knowledge and Data Engineering*, *25*(8), 1719–1731.

Breck, E., Choi, Y., & Cardie, C. (2007). Identifying expressions of opinion in context. *International Joint Conferences on Artificial Intelligence*, *7*, 2683–2688.

Cogburn, D. L., & Espinoza-Vasquez, F. K. (2011). From networked nominee to networked nation: examining the impact of Web 2.0 and social media on political participation and civic engagement in the 2008 Obama campaign. *Journal of Political Marketing*, *10*(1–2), 189–213.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Dang, Y., Zhang, Y., & Chen, H. (2010). A lexicon-enhanced method for sentiment classification: an experiment on online product reviews. *IEEE Intelligent Systems*, *25*(4), 46–53.

Euzenat, J., & Shvaiko, P. (2007). . *Ontology matching* (Vol. 18) Heidelberg, Germany: Springer.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, *3*, 1289–1305.

Galam, S., Gefen, Y., & Shapir, Y. (1982). Sociophysics: a new approach of sociological collective behaviour. I. Mean-behaviour description of a strike. *Journal of Mathematical Sociology*, *9*(1), 1–13.

Galam, S., & Moscovici, S. (1991). Towards a theory of collective phenomena: consensus and attitude changes in groups. *European Journal of Social Psychology*, *21*(1), 49–74.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, *61*, 115–125.

Graham, T., Broersma, M., Hazelhoff, K., & van't Haar, G. (2013). Between broadcasting political messages and interacting with voters: the use of Twitter during the 2010 UK general election campaign. *Information, Communication & Society*, *16*(5), 692–716.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1–3), 389–422.
Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., & Kaymak, U. (2013). Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing* (pp. 703–710).
Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, *60*(11), 2169–2188.
Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, *22*(2), 110–125.
KOMORAN (2003). http://www.shineware.co.kr.
Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: the good the bad and the omg!. *ICWSM*, *11*, 538–541.
Laboreiro, G., Sarmento, L., Teixeira, J., & Oliveira, E. (2010). Tokenizing micro-blogging messages using a text classification approach. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data* (pp. 81–88). ACM.
Larsson, A. O., & Moe, H. (2012). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society*, *14*(5), 729–747.
Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.
Matsumoto, S., Takamura, H., & Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency sub-trees. In *Advances in Knowledge Discovery and Data Mining*. pp. 301–311. Heidelberg, Germany: Springer.
McCallum, A. K. (2002). MALLET: a machine learning for language toolkit. Available at http://mallet.cs.umass.edu.
Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 53–71.
Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: an empirical comparison between SVM and ANN. *Expert Systems with Applications*, *40*(2), 621–633.
Na, S. H., Lee, Y., Nam, S. H., & Lee, J. H. (2009). Improving opinion retrieval based on query-specific sentiment lexicon. In *ECIR 09*. pp. 734–738. Berlin, Germany: Springer.
Nooralahzadeh, F., Arunachalam, V., & Chiru, C. (2013). 2012 Presidential Elections on Twitter—an analysis of how the US and French Election were teflected in tweets. In *2013 19th International Conference on Control Systems and Computer Science (CSCS)* (pp. 240–246). IEEE.
Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA)* (pp. 19–21).
Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing* (Vol. 10) (pp. 79–86). Association for Computational Linguistics.
Pelikan, M. (2005). *Bayesian optimization algorithm*. pp. 31–48. *Hierarchical Bayesian optimization algorithm* (Vol. 170) Heidelberg, Berlin: Springer.
Prasad, S. S., Kumar, J., Prabhakar, D. K., & Pal, S. (2015). Sentiment classification: an approach for Indian language tweets using decision tree. In *Mining intelligence and knowledge exploration*. pp. 656–663. Springer International Publishing.
Saleh, M. R., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. (2011). Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications*, *38*(12), 14799–14804.
Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 267–288.
Tsai, A. C. R., Wu, C. E., Tsai, R. T. H., & Hsu, J. Y. J. (2013). Building a concept-level sentiment dictionary based on commonsense knowledge. *IEEE Intelligent Systems*, *28*(2), 22–30.
Wegrzyn-Wolska, K., & Bougueroua, L. (2012). Tweets mining for French presidential election. In *2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN)* (pp. 138–143). IEEE.
Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, *35*(3), 399–433.
Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, *181*(6), 1138–1152.
Zhang, L., & Liu, B. (2011). Identifying noun product features that imply opinions. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Vol. 2) (pp. 575–580). Association for Computational Linguistics.
Zhang, D., Ma, J., Yi, J., Niu, X., & Xu, X. (2015). An ensemble method for unbalanced sentiment classification. In *2015 11th International Conference on Natural Computation (ICNC)* (pp. 440–445). IEEE.
Zhang, Z., Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems with Applications*, *38*(6), 7674–7682.