1-2020

Optimal feature selection for learning-based algorithms for sentiment classification

Zhaoxia WANG
Singapore Management University, zxwang@smu.edu.sg

Zhiping LIN
Nanyang Technological University

Citation

WANG, Zhaoxia and LIN, Zhiping. Optimal feature selection for learning-based algorithms for sentiment classification. (2020). *Cognitive Computation*. 12, (1), 238-248.
Available at: https://ink.library.smu.edu.sg/sis_research/5887

Published in Cognitive Computation, Volume 12, Issue 1, Pages 238-248
https://doi.org/10.1007/s12559-019-09669-5

Optimal Feature Selection for Learning-Based Algorithms for Sentiment Classification

Zhaoxia Wang [1,2,3] **· Zhiping Lin** [4]

Abstract

Sentiment classification is an important branch of cognitive computation—thus the further studies of properties of sentiment analysis is important. Sentiment classification on text data has been an active topic for the last two decades and learning-based methods are very popular and widely used in various applications. For learning-based methods, a lot of enhanced technical strategies have been used to improve the performance of the methods. Feature selection is one of these strategies and it has been studied by many researchers. However, an existing unsolved difficult problem is the choice of a suitable number of features for obtaining the best sentiment classification performance of the learning-based methods. Therefore, we investigate the relationship between the number of features selected and the sentiment classification performance of the learning-based methods. A new method for the selection of a suitable number of features is proposed in which the Chi Square feature selection algorithm is employed and the features are selected using a preset score threshold. It is discovered that there is a relationship between the logarithm of the number of features selected and the sentiment classification performance of the learning-based method, and it is also found that this relationship is independent of the learning-based method involved. The new findings in this research indicate that it is always possible for researchers to select the appropriate number of features for learning-based methods to obtain the best sentiment classification performance. This can guide researchers to select the proper features for optimizing the performance of learning-based algorithms. (A preliminary version of this paper received a Best Paper Award at the International Conference on Extreme Learning Machines 2018.)

Keywords Machine learning · Feature selection · Optimal feature selection · Relationship analysis · Sentiment classification · Social media · Text analysis

Introduction

Sentiment analysis is an important branch of natural language processing (NLP). NLP is one of the important topics of cognitive computation. Thus the further studies of the properties of sentiment analysis methods are important research for cognitive computation [1]. Sentiment analysis has been used to analyze the social media which has become very popular nowadays [1–3]. Feng et al. mentioned that analyzing human sentiments is a critical problem in cognitive computing [2]. They believe that one fundamental task of sentiment analysis is to infer the sentiment polarity or emotion category of subjective text, such as microblogs and they proposed a novel method for detecting emotions in the short text of microblogs [2]. Wang and Wu conducted sentiment analysis on Twitter messages and they also claimed that predicting people's sentiments from their posts is a critical problem in cognitive computing [3].

✉ Zhaoxia Wang
zxwang@smu.edu.sg; zhxwang@nuist.edu.cn

Zhiping Lin
EZPLin@ntu.edu.sg

[1] School of Information Systems, Singapore Management University (SMU), 80 Stamford Road, Singapore 178902, Singapore

[2] Nanjing University of Information Science and Technology (NUIST), No. 219, Ningliu Road, Nanjing, Jiangsu 210044, China

[3] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Singapore

[4] School of Electrical and Electronic Engineering (EEE), Nanyang Technological University, Singapore 639798, Singapore

More and more people prefer to post their opinions or sentiments on different topics such as politics, products, or movies through social media platforms. These data contain a large amount of up-to-date information, which can be utilized by different organizations for different purposes [4, 5]. For example, tweets had been collected for analyzing the opinions of Twitter users for predicting the outcome of 2016 US Presidential Election [6]. Governments may be interested in knowing how people think about their political decisions and companies may want to know whether their customers are satisfied with their services or products through social media sentiment analysis. Sentiment analysis can also be applied to solve problems in different domains, such as healthcare [5] and business marketing [7, 8].

Sentiment classification is a subfield of sentiment analysis. It refers to the classification of opinion comments into positive, negative, or neutral categories. There are basically two types of approaches to do sentiment analysis: non-learning-based methods, such as lexicon-based approaches, and learning-based methods, such as Naïve Bayes (NB) and Maximum Entropy (MaxEnt) [9, 10]. Lexicon-based approaches classify the sentence by checking the meaning of each word or phrase using existing lexicons while machine learning-based approaches use statistical-based or non-statistical-based inference for the classification [9, 10]. In addition to the basic non-learning-based and learning-based methods, hybrid methods which combine the two basic methods have also been proposed for sentiment analysis [11–20].

In learning-based approaches, each word or phrase represents a feature. During the training process, these features are used to train the learning-based methods. It is obvious that not all of these features are useful in the actual analysis. There are some noisy features that may affect the results. Thus, selecting the appropriate number of useful and valuable features is important for the learning-based methods to perform well. If the number of features selected is too large, it will make the training time-consuming and too expensive computationally. If the number of features used is too small, the useful features may not be selected, and it may affect the analysis results. Thus, optimal feature selection is necessary for a learning-based method to achieve a satisfactory sentiment classification performance and it may also save computational resources. However, currently, there is no solution to determine the size of the subset of the entire set of features to be used to obtain the best sentiment classification performance. It is still a challenging problem for researchers.

In this paper, we design a new method to investigate the relationship between the number of features used and the sentiment classification performance of the learning-based algorithms. Different learning-based methods, such as NB, MaxEnt, support vector machine (SVM), and extreme learning machine (ELM) are investigated in this study.

The results demonstrate that feature selection has a significant effect on the learning process of the learning-based algorithms. A relationship between the number of features selected and the sentiment classification performance of the algorithms is discovered through this research work. The results show that the relationship between the number of features used and the performance of the algorithms is consistent for different learning-based algorithms which means that this relationship is independent of the algorithms used. Specifically, this paper presents a new and very clear relationship between the performance of the learning-based algorithms and the logarithm of the number of features to be selected.

The novelty and importance of this paper are summarized as follows:

1. All the existing works have emphasized that feature selection is very important for sentiment classification. A few works have discussions on selecting an appropriate number of features to obtain the best performance. However, none of these works discuss the exact relationship between feature selection and the classification performance of the methods.
2. In these existing works, there is no discussion on how the accuracy of the machine learning algorithms changes with different number of features selected. It is still an unsolved problem and un-answered question on how to select a suitable number of features to be used to obtain the best performance for analyzing different sizes of data.
3. This research answers the question and solves the problem by presenting a novel methodology. This is the first time that a very clear relationship between the performance of the learning-based algorithms and the number of features to be selected is discovered. The relationship discovered in this research has never been disclosed by any existing work.

The rest of the paper is organized as follows: “Related Work” section discusses the related work. The proposed methodology is presented in “Proposed Methodology” section. “Dataset Collection and Preparation” section describes the datasets used by this research. The results are shown and discussed in “Results and Discussion” section which is followed by further analysis on larger datasets in “Further Analysis on Larger Datasets” section. “Conclusion and Future Work” section presents the conclusion and future work.

Related Work

Sentiment analysis has gained the interest of many researchers [21–23]. For example, sentiment analysis on product reviews posted online can reveal how different age groups act on the purchasing of online products [24]. It can help merchants understand how the customers feel about their products and how they can improve the products involved [25]. Likewise, public moods as uncovered by sentiment analysis have been found to correlate with market movements. For example, Malandri et al. demonstrated that applying sentiment analysis as part of their approach led to greater expected returns [26]. Dashtipour et al. presented a state-of-the-art review on multilingual sentiment analysis, highlighting how different preprocessing techniques and different methods across languages have their own advantages and disadvantages [21].

Machine learning approaches are often preferred as compared to lexicon-based approaches [22]. Narayanan et al. proposed an improved NB model for fast and accurate sentiment classification which obtained quite good performance: the accuracy hits 88.8% on the movie review data [22]. Wang et al. investigated various enhancement strategies such as emoticon handling and negation handling to improve on the different machine learning algorithms and the performance of the methods increases with the assistance of the enhancing techniques [23].

Hybrid methods which combine the non-learning-based method and learning-based methods are also used in sentiment analysis [11–20]. Cambria et al. proposed a novel method which simultaneously trains a sentiment classifier and adapts an existing sentiment lexicon to the target domain. Their method significantly improved the sentiment classification performance for a variety of domains by means of improving the quality of the sentiment lexicons [11]. SenticNet 5, another piece of work by Cambria et al., combines commonsense knowledge representation with deep learning to achieve good performance [12]. Mondal et al. combined a learning-based method with a domain-based knowledge lexicon to extract semantic relations in healthcare domains, in which the sentiment of the medical concepts involved were considered as positive or negative. The resulting approach is a concept clustering method that identifies the semantic relations of concepts to enhance clinical decision-making in healthcare systems [13]. Lauren et al. developed an ELM-based word embedding approach for NLP; their results demonstrated the merits of the methods for sentiment analysis and sequence labeling [14]. Li et al. incorporated the prior sentiment information consisting of document level sentiment and word level sentiment to enhance the performance of the methods involved. Different combinations of the prior information (different lexicons with different document level sentiment) were tested on different machine learning algorithms. Their results showed that the sentiment classification performance of the method involved can be improved with the hybrid enhancement strategies [15]. Kolchyna et al. used lexicon to measure the score of each sentence [16]. Then these scores were fed into learning machines as additional features, leading to an increase in accuracy. Likewise, Zhang et al. used a lexicon-based method to automatically create labels that could serve as training data, thus allowing machine learning to be performed in an unsupervised manner. Cambria et al. considered the difference between human intelligence and traditional artificial intelligence (AI) and exploited common sense knowledge to perform reasoning as humans do [18]. They also proposed a unique multi-disciplinary method for sentiment analysis [19]. Their technique organizes the features and other common knowledge in a better way in the vector space. Wang et al. proposed an intelligent sensing mechanism using an adaptive learning method which combines non-learning-based and learning-based method in a unique way to enhance the performance of the sentiment analysis of text data [20].

Before applying lexicon-based or machine learning-based methods, various algorithms are employed for preprocessing [9, 12, 18, 23, 27, 28]. Feature selection is one of these algorithms and different feature selection algorithms such as Chi Square, term frequency inverse document frequency, and information gain are used by many researchers [28–30]. All these existing works have emphasized that feature selection is very important for sentiment classification. Not only would it improve accuracy, but computational time can also be reduced [31, 32]. However, only a few works have discussed how an appropriate number of features should be selected to obtain the best sentiment classification performance.

Some new feature selection algorithms have recently been developed [31, 32]. Al-Radaideh et al. proposed a new feature selection method based on rough sets for Arabic sentiment analysis. They have built a rough engine for sentiment classification based on the proposed feature selection method. The rough set model was tested and compared with other machine learning algorithms. The result showed that the proposed model in [31] outperforms other machine learning algorithms, but there is no discussion on how the accuracy of the machine learning algorithms changes with different number of features selected [31]. Narayanan et al. plotted a graph of accuracy vs number of features and found that choosing the top 32000 features will produce the best results [22]. However, there is also no discussion on the relationship between the number of features to be selected and the performance of the algorithm in their paper. The impact of feature selection has also been discussed by Prusa et al. [32]. Ten feature ranking methods including Chi Square (CS) and Gini-Index (GI) were compared using 4 different machine learning algorithms with different numbers of features selected. The results across different combinations were discussed but no relationship between the number of features and accuracy was found in their experiment.

Although almost every researcher who engages in text analysis knows that feature selection is important for learning-based methods, it is still an unsolved problem and un-answered question on how to select a suitable number of features to be used to obtain the best sentiment classification performance. None of these existing works discuss an explicit relationship between the number of features selected and the sentiment classification performance of the methods.

In this work, we investigate the relationship between the number of features used and the sentiment classification performance of the learning-based methods. Different learning-based methods and different sizes of data are used, and the Chi Square method is employed to perform feature selection as it is one of the top ranking methods used by previous researchers [32]. We present different score thresholds for the selection of the number of features in which a new measure, log(no_of_features), is proposed to be used for the relationship analysis. The sentiment classification performances of different machine learning methods with different sets of features selected are compared and evaluated.

Proposed Methodology

A new methodology is proposed in this paper for analyzing the relationship between the selection of a suitable number of features and the performance of the learning-based methods. The detailed methods are described in this section which include data preprocessing, implementation of learning-based methods, and variable mathematical transformation for feature selection.

Data Preprocessing

Some basic preprocessing techniques are used before analyzing the data: (1) decode html into the encoded strings, (2) convert all upper case letters to lower case letters, (3) convert links to “url,” (4) convert “@username” to “user,” (5) remove all the punctuations, (6) convert “#word” to “word,” (7) remove repeated letters, (8) add “NEG” to the words that is after the negate words, and (9) remove stop words that add no meaning to the content.

Implementation of Learning-Based Methods

Machine learning methods have been effectively applied to sentiment analysis. Among these methods, NB, MaxEnt, and SVM are relatively well-known [22, 33]. Besides these algorithms, ELM, a recently developed machine learning algorithm, has also been applied to many areas including sentiment analysis. Hence, NB, MaxEnt, SVM, and ELM are selected as base methods in this research.

The NB algorithm is a probabilistic method based on the Bayes Theorem and Maximum Entropy (MaxEnt) is another probability-based method [33]. MaxEnt is a technique that models the probability distribution from the training data. The principle of MaxEnt states that the distribution should be as uniform as possible when there is no observation. In the field of text sentiment classification, MaxEnt uses word features in the labeled documents as constraints to optimize the conditional distribution.

Proposed by Cortes and Vapnik, SVM is a kind of support-vector neural network. The general idea is to map the input vectors onto some high dimensional feature space so that the input vectors can be separated by a hyperplane in that space [34].

ELM, first proposed by Huang [35], is based on the feedforward neural network. Traditionally, the weights and the biases in each layer must be tuned in order to obtain the best results. However, it is not necessary to tune these parameters if they are randomly generated, which is the key idea in ELM. In this case, this feedforward neural network can be considered as a linear system. Then, the output weights can be determined by inversed operations carried out on the hidden layer output matrices.

Variable Mathematical Transformation Setting for Feature Selection

Feature selection refers to selecting the subset of the entire set of features. In sentiment classification, the purpose of feature selection is to make the methods more efficient and accurate [36, 37]. There are a lot of feature selection methods, among which the Chi Square feature selection method is a very popular and efficient one. In this paper, the Chi Square method [23, 37] is employed for feature selection.

After scoring all the individual features using the Chi Square method, we use a score-based method to select the features. The features with higher scores are selected to be used by the learning-based methods. Different score thresholds are used to select the top number of features. In order to test and estimate the exact relationship between the number of features used and the sentiment classification performance of the learning-based method, we increase the number of features selected from 10 to 10^4 and propose a new measure, *Fn*, which is obtained by computing the logarithm of the number of features to the base 10:

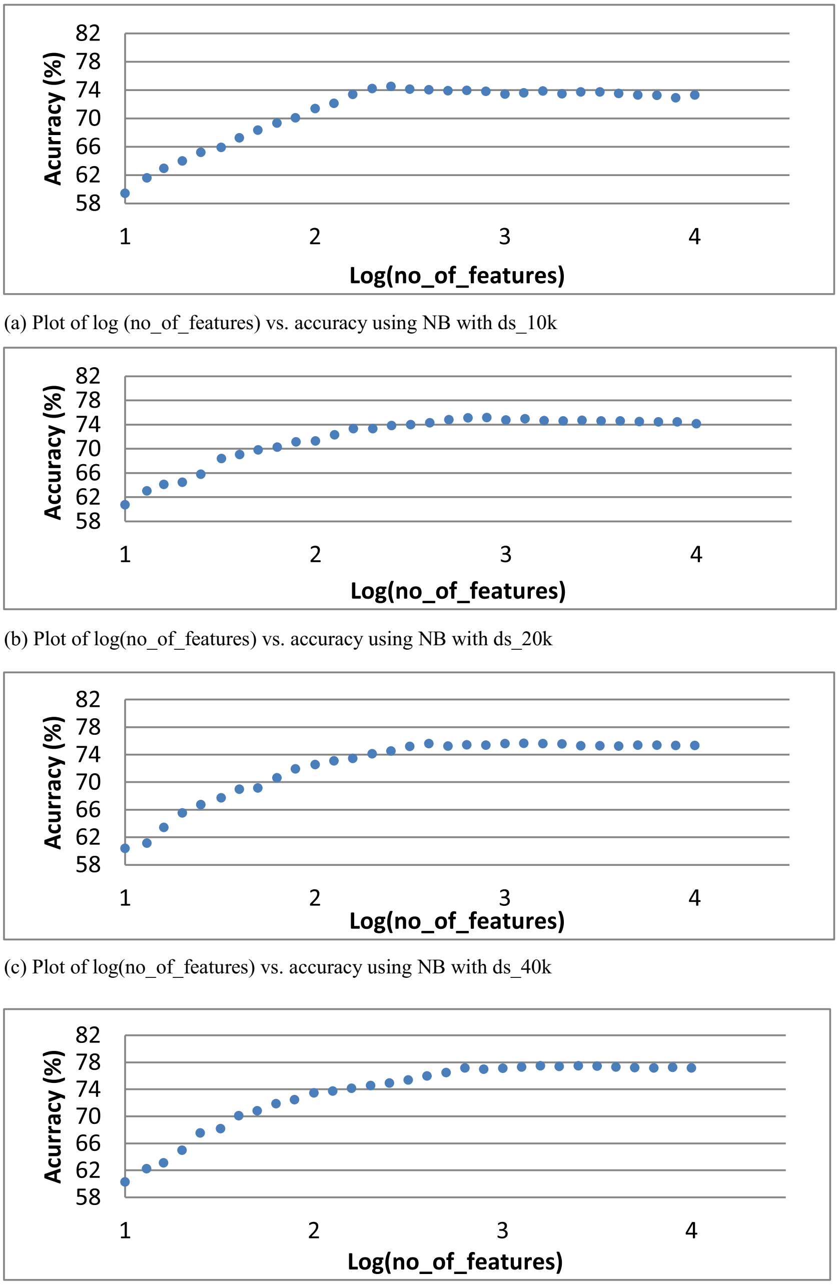
$$\mathrm{Fn} = \log(\mathrm{no\_of\_features}) \quad (1)$$

where no _ of _ features represents the number of features selected to be used by the methods. After introducing such a mathematical transformation, the relationship between the number of features used and the performance of the learning-based method is discovered. The detailed results are shown in “Results and Discussion” section by using the datasets described in “Dataset Collection and Preparation” section.

Dataset Collection and Preparation

In this paper, we used two publicly available datasets. The first (named dataset A) was downloaded from the “Twitter-sentiment-analyzer” website, which provides 1.6 million pre-classified tweets [38]. We extracted different sizes of data, including 10k, 20k, 40k, and 80k tweets from the dataset downloaded and named them respectively as ds_A10k, ds_ A20k, ds_ A40k, and ds_ A80k. The second dataset (named dataset B) was from “Twitter-sentiment-analysis2” [39]. Similar to dataset A, we extracted different sizes of data for the comparison analysis. We perform both automated and

Fig. 1 The results obtained from analyzing dataset A in different sizes using NB. **a** Plot of log (no_of_features) vs. accuracy using NB with ds_10k. **b** Plot of log(no_of_features) vs. accuracy using NB with ds_20k. **c** Plot of log(no_of_features) vs. accuracy using NB with ds_40k. **d** Plot of log(no_of_features) vs. accuracy using NB with ds_80k



(a) Plot of log (no_of_features) vs. accuracy using NB with ds_10k

(b) Plot of log(no_of_features) vs. accuracy using NB with ds_20k

(c) Plot of log(no_of_features) vs. accuracy using NB with ds_40k

(d) Plot of log(no_of_features) vs. accuracy using NB with ds_80k

manual data cleansing processes to obtain ground truth datasets.

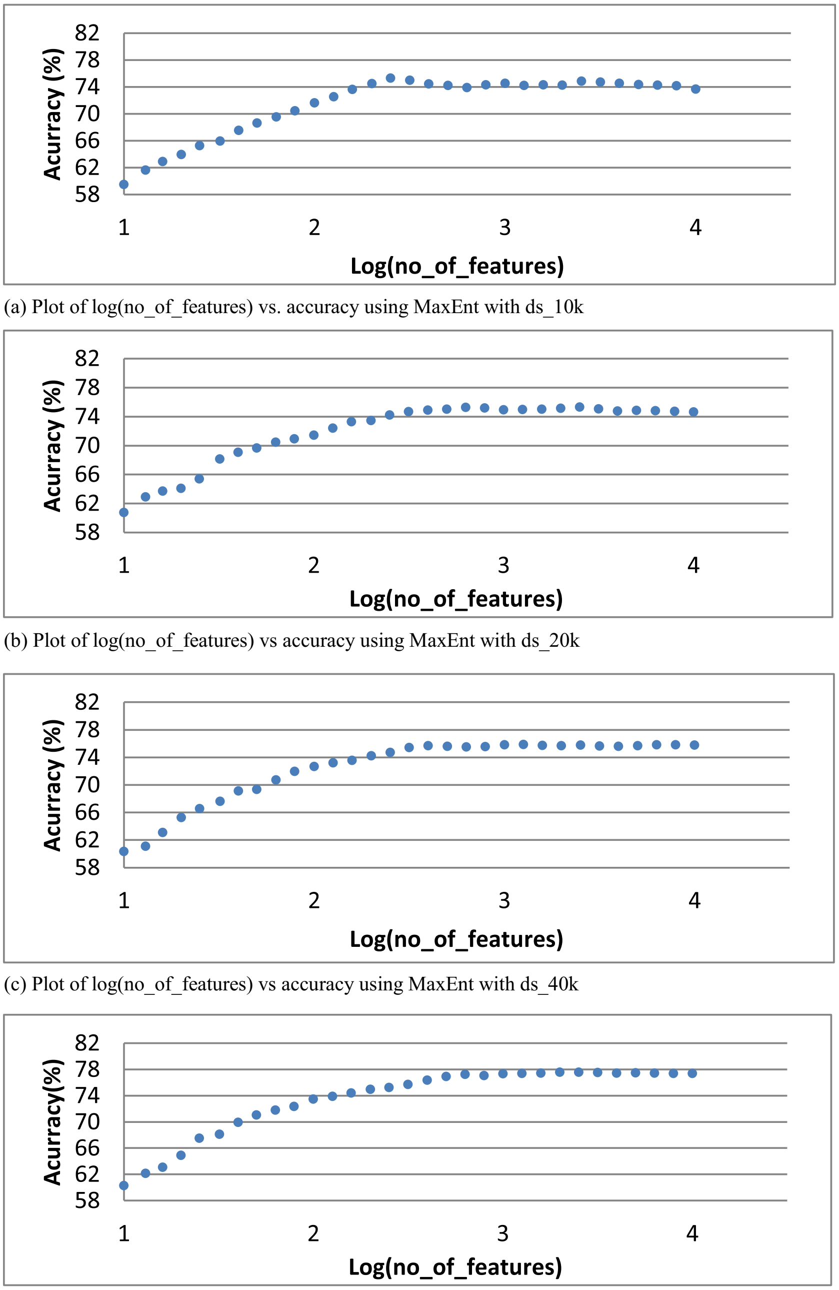
All the datasets are balanced: 50% of the data are positive and the other 50% are negative. We split each dataset into training data, which consist of three quarters of the data, and testing data, which are the remaining one quarter.

Results and Discussion

For dataset A, four datasets in different sizes, 10k, 20k, 40k, and 80k are analyzed and tested through the feature selection method proposed above. Four learning-based methods, NB, MaxEnt, SVM, and ELM algorithms, are used. Figures 1, 2, 3, and 4 are the results obtained from analyzing the four datasets using NB, MaxEnt, SVM, and ELM respectively. The results are presented as plots of log(no_of_features) (*x*-axis) versus accuracy of the learning-based method (*y*-axis). In Figs. 1, 2, 3, and 4, the 4 sub-figures, a–d, correspond to the results obtained from the 4 different datasets of 10k, 20k, 40k, and 80k respectively.

From the results obtained by NB, MaxEnt, and SVM (Figs. 1, 2, and 3), we can see that the sentiment classification accuracy

Fig. 2 The results obtained on analyzing dataset A in different sizes using MaxEnt. **a** Plot of log(no_of_features) vs. accuracy using MaxEnt with ds_10k. **b** Plot of log(no_of_features) vs accuracy using MaxEnt with ds_20k. **c** Plot of log(no_of_features) vs accuracy using MaxEnt with ds_40k. **d** Plot of log(no_of_features) vs accuracy using MaxEnt with ds_80k



(a) Plot of log(no_of_features) vs. accuracy using MaxEnt with ds_10k

(b) Plot of log(no_of_features) vs accuracy using MaxEnt with ds_20k

(c) Plot of log(no_of_features) vs accuracy using MaxEnt with ds_40k

(d) Plot of log(no_of_features) vs accuracy using MaxEnt with ds_80k

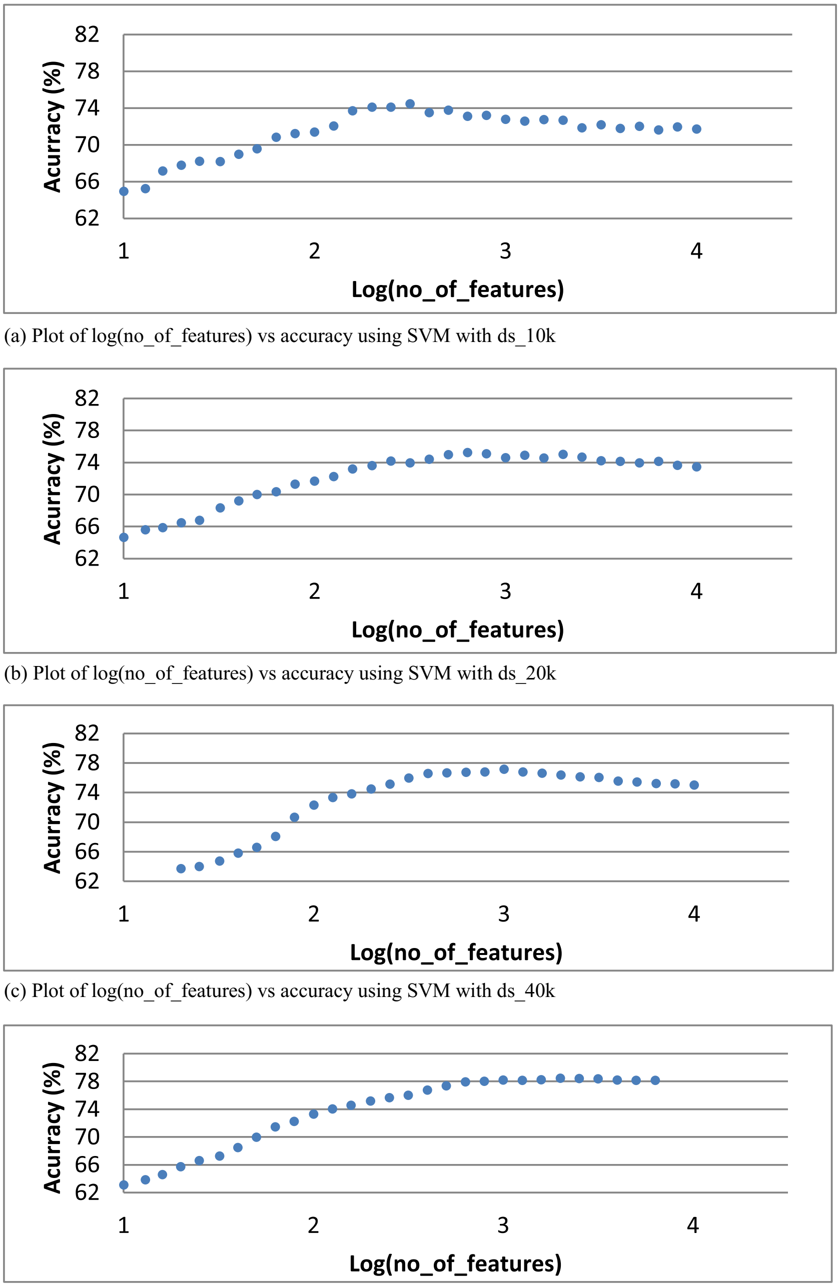
increases proportionally to the increase in the number of features until the accuracy hits a maximum (optimal) point.

Figures 4 a and b show the results of analyzing the two datasets in smaller sizes (10k and 20k) by using the ELM method and Figs. 4 c and d show the results of analyzing the two datasets in larger sizes (40k and 80k) by using the same ELM method. It is observed that the plots of Figs. 4 a and b differ from that of Figs. 4 c and d. For Figs. 4 a and b, the accuracy increases generally with the increase in the log(no_of_features) but not as uniformly proportional as that in Figs. 4 c and d.

In Figs. 4 a and b, the accuracy oscillates somewhat with the increase in the log(no_of_features). This may be because the size of the datasets used to train the ELM is too small to enable a stable performance of the algorithm. The results obtained from using the ELM method on the other 2 larger datasets (Figs. 4 c and d) show a clear stable relationship between log(no_of_features) and accuracy of the algorithm.

This result indicates that ELM is very suitable for analyzing large datasets, which is consistent with previous research by Liu et al. [40], who indicated that the

Fig. 3 The results obtained from analyzing dataset A in different sizes using SVM. **a** Plot of log(no_of_features) vs accuracy using SVM with ds_10k. **b** Plot of log(no_of_features) vs accuracy using SVM with ds_20k. **c** Plot of log(no_of_features) vs accuracy using SVM with ds_40k. **d** Plot of log(no_of_features) vs accuracy using SVM with ds_80k



(a) Plot of log(no_of_features) vs accuracy using SVM with ds_10k

(b) Plot of log(no_of_features) vs accuracy using SVM with ds_20k

(c) Plot of log(no_of_features) vs accuracy using SVM with ds_40k

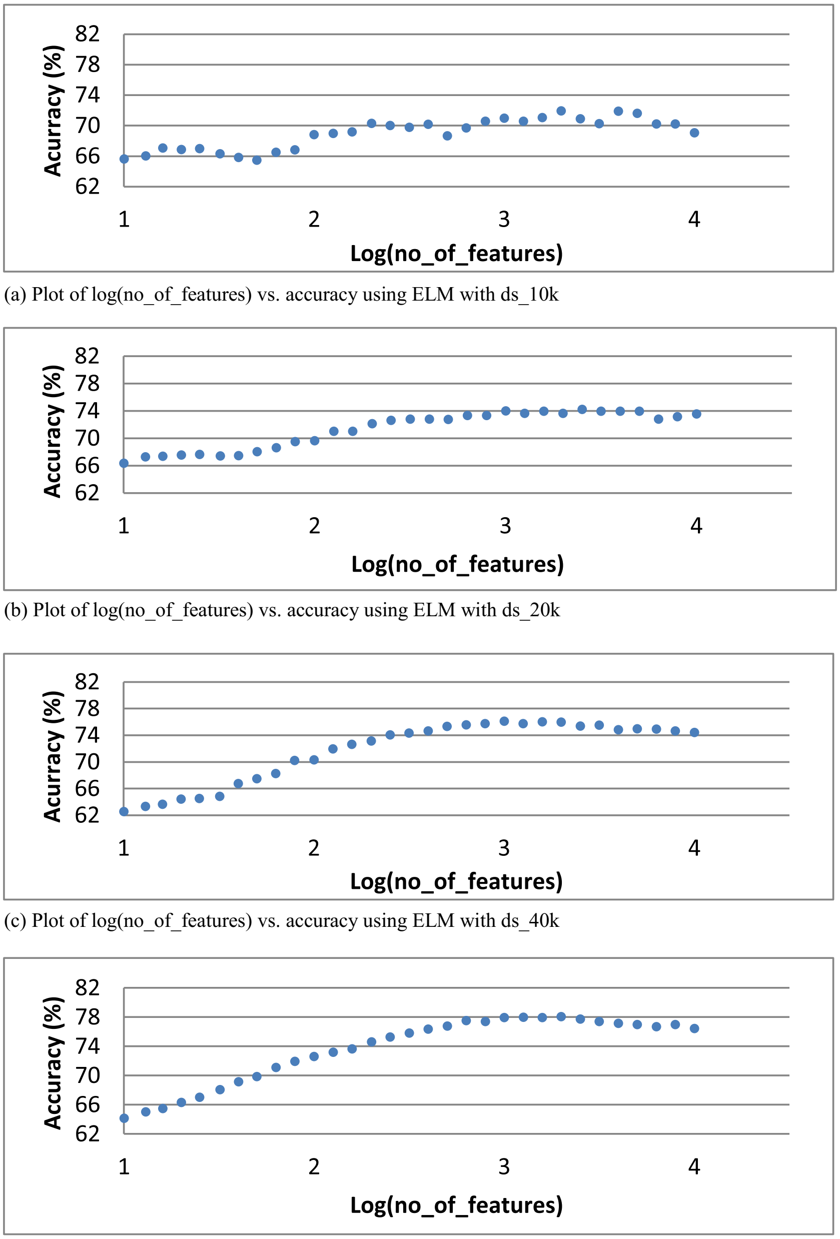
(d) Plot of log(no_of_features) vs accuracy using SVM with ds_80k

generalization ability of ELM is poor when the size of the dataset is small but has the potential to yield good generalization behavior when the size of the dataset becomes large.

Analyzing all the results obtained in this research as shown in the 16 plots of Figs. 1, 2, 3, and 4, the overall trend of the relationship between the sentiment classification accuracy of the method and log(no_of_features) is clearly shown: with the increase in the log(no_of_features), the accuracy increases before it hits a maximum or optimal point, especially for larger datasets. After it hits the maximum or optimal point, where the value of log(no_of_features) is around 3 (roughly between 2.5 and 3.5), corresponding to about 1,000 features, the performance of the algorithm will not continue to increase, instead, it stays constant or decreases slightly.

Similar results are obtained when analyzing dataset B for all the four learning-based methods as well: there is a clear relationship between log(no_of_features) and the accuracy of the learning-based algorithm. There is also an optimal point for the log(no_of_features) which means that there exists an optimal number of features.

Fig. 4 The results obtained from analyzing Dataset A in different sizes using ELM. **a** Plot of log(no_of_features) vs. accuracy using ELM with ds_10k. **b** Plot of log(no_of_features) vs. accuracy using ELM with ds_20k. **c** Plot of log(no_of_features) vs. accuracy using ELM with ds_40k, **d** Plot of log(no_of_features) vs. accuracy using ELM with ds_80k



(a) Plot of log(no_of_features) vs. accuracy using ELM with ds_10k

(b) Plot of log(no_of_features) vs. accuracy using ELM with ds_20k

(c) Plot of log(no_of_features) vs. accuracy using ELM with ds_40k

(d) Plot of log(no_of_features) vs. accuracy using ELM with ds_80k

The relationship is also independent of the learning-based methods involved. To save space, we do not present the sentiment classification results on dataset B but they are available upon request.

Table 1 *R*-squared values for trendline analysis of the relationship between log(no_of_feature) and accuracy for different learning-based methods

	R-squared values			
	NB	MaxEnt	SVM	ELM
2nd order polynomial	0.9852	0.9864	0.9946	0.9913
3rd order polynomial	0.9955	0.9968	0.9954	0.9960
4th order polynomial	0.9963	0.9971	0.9977	0.9981

The reason behind this general pattern can be explained as follows: at the start of the regime, the features used are all useful. When the number of useful features increases, the sentiment classification accuracy of the learning-based methods also increases. At the maximum point, all the useful features have already been selected. Beyond this point, features are more likely to be “not important” or “noisy” features. Because of these noisy features, the performance of the learning-based methods does not improve further, and the accuracy of the methods will remain the same or even start to decrease as shown in Figs. 1, 2, 3, and 4.

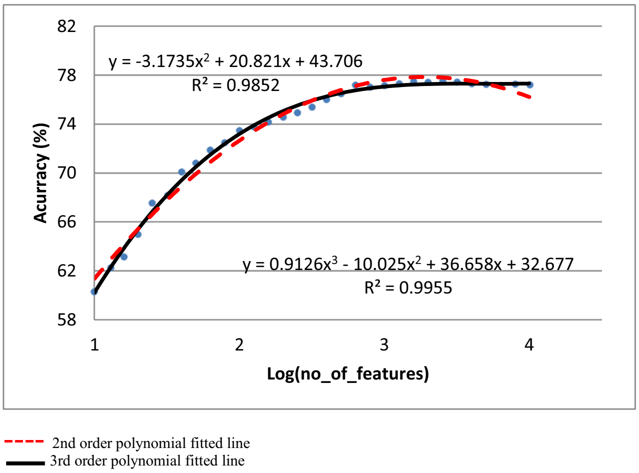


Fig. 5 Trendline analysis on the relationship between log(no_of_features) and accuracy using NB (80k). ---- 2nd order polynomial fitted line. ▬ 3rd order polynomial fitted line

Further Analysis on Larger Datasets

Encouraged by the findings presented in "Results and Discussion" section, we carried out further analysis by trying to fit polynomial equations on the relationship between log(no_of_features) and the accuracy of the learning-based methods for large datasets (80k). Specifically, an *n*th order polynomial has the following form:

$$y = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + b \tag{2}$$

where *y* represents the accuracy of the learning-based methods and *x* represents log(no_of_features).

We use 2nd, 3rd, and 4th order polynomials to fit the data and compare the differences in terms of *R*-squared values [41]:

$$R\text{–squared} = \text{explained variance}/\text{outcome variance} \tag{3}$$

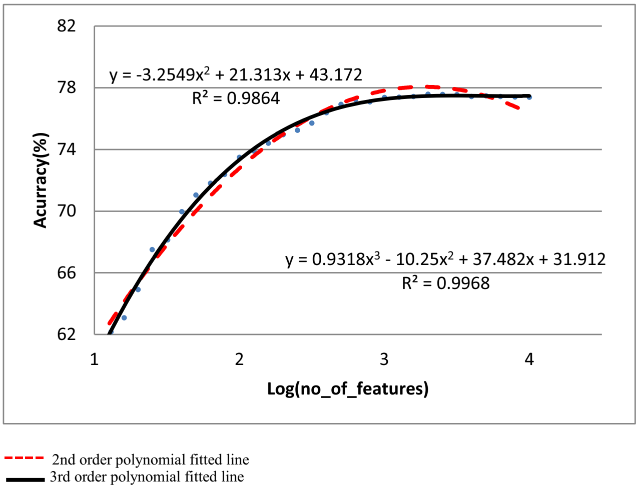


Fig. 6 Trendline analysis on the relationship between log(no_of_features) and accuracy using MaxEnt (80k). ---- 2nd order polynomial fitted line. ▬ 3rd order polynomial fitted line

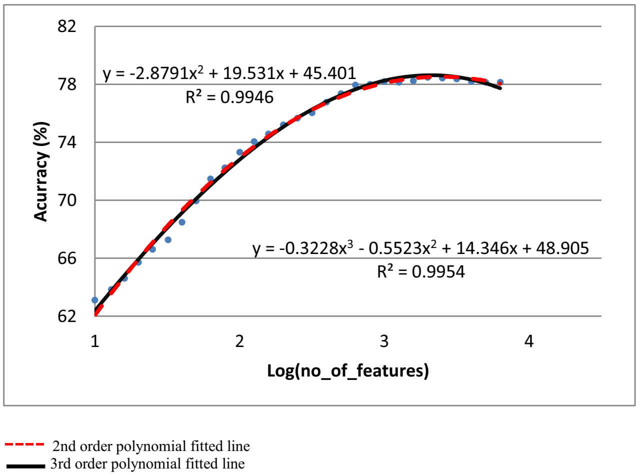


Fig. 7 Trendline analysis on the relationship between log(no_of_features) and accuracy using SVM (80k). ---- 2nd order polynomial fitted line. ▬ 3rd order polynomial fitted line

R-squared is a statistical measure with value between 0 and 1. It is used to calculate how the data are close to the trendline fitted. Generally, the higher the *R*-squared value, the better the model fits.

It is discovered that the *R*-squared values are always greater than 0.98 whether 2nd, 3rd, or 4th order polynomials are used. Comparing the 2nd, 3rd, and 4th order polynomials, it is found that the differences are very small as shown in Table 1. The results are also shown in Figs. 5, 6, 7, and 8 for the four learning-based methods for the 2nd and 3rd order polynomials.

Based on the results shown in Table 1 and in Figs. 5, 6, 7, and 8, it is observed that the 2nd order polynomial or quadratic function is good enough to represent the relationship:

$$y = a_2 x^2 + a_1 x + b \tag{4}$$

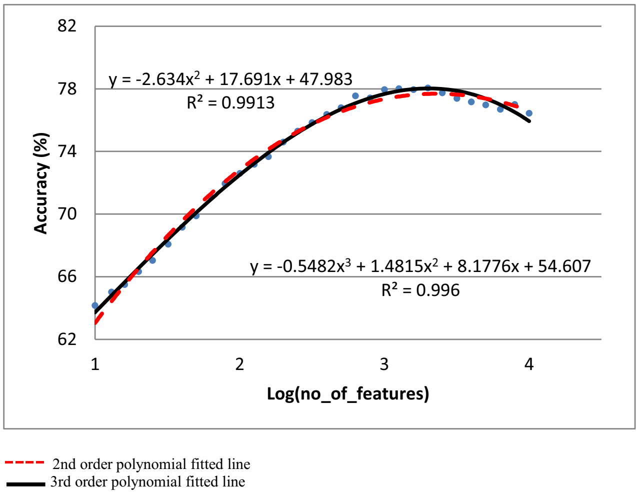


Fig. 8 Trendline analysis on the relationship between log(no_of_features) and accuracy using ELM (80k). ---- 2nd order polynomial fitted line. ▬ 3rd order polynomial fitted line

This again demonstrates that for learning-based methods, consistent relationships, in this case, between log(no_of_features) and accuracy are obtainable for large datasets.

Conclusion and Future Work

In this paper, an optimal feature selection method for sentiment classification is proposed and investigated. The sentiment classification performance of different machine learning methods in terms of accuracy is studied using the proposed feature selection methodology. Various numbers of features are selected based on the feature scores for learning-based method on sentiment analysis to investigate whether there is a relationship between the number of features to be selected and the sentiment classification performance of the learning-based algorithms. A feature score threshold is used for feature selection. A relationship between the sentiment classification accuracy of the learning algorithms and the logarithm of the number of features selected is discovered by this research. This new finding can guide researchers to select the proper features for optimizing the sentiment classification performance of learning-based methods. By optimizing the features used for the learning-based algorithms involved, better sentiment classification performance of the learning-based method could be achieved. This also demonstrates that researchers can select the proper number of features to obtain the optimized performance of the machine learning-based methods.

Our ongoing work also includes using deep learning for sentiment analysis. We will also investigate why the sentiment classification performance of ELM is not stable for analyzing datasets of smaller sizes. Also, we would like to investigate the performance of the learning-based method for multi-level sentiment classification. Lastly, carrying on studies on improving the sentiment classification performance of hybrid methods, which combine the non-learning-based and learning-based methods, is also planned in our future work.

Acknowledgments The authors would like to thank Dr. Ho Seng Beng, Dr. Quek Boon Kiat, and the team of A*STAR AI program for their discussion and help. The authors would also like to thank the intern students from NTU and SUTD for the assistance in this research.

Compliance with Ethical Standards

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Asgarian E, Kahani M, Sharifi S. The impact of sentiment features on the sentiment polarity classification in Persian reviews. Cognit Comput. 2018;10(1):117–35.
2. Feng S, Wang Y, Song K, Wang D, Yu G. Detecting multiple coexisting emotions in microblogs with convolutional neural networks. Cognit Comput. 2018;10(1):136–55.
3. Yang H, Wu CLC. Sentiment discovery of social messages using self-organizing maps. Cognit Comput. 2018;10(6):1152–66.
4. Dashtipour K, Gogate M, Adeel A, Ieracitano C, Hussain A. Exploiting deep learning for Persian sentiment analysis. Int Conf Brain Inspired Cognit Syst. 2018:597–604.
5. Cambria E, Hussain A, Durrani T, Havasi C, Eckl C, Munro J. Sentic computing for patient centered applications. Proc IEEE ICSP. 2010:1279–82.
6. Bovet A, Morone F, Makse HA. Validation of Twitter opinion trends with national polling aggregates : Hillary Clinton vs Donald Trump. Sci Rep. 2018;8(1):8673.
7. Wang Z, Tong JC, Xin X, Chin HC. Anomaly detection through enhanced sentiment analysis on social media data. In: 2014 IEEE 6th international conference on cloud computing technology and science; 2014. p. 917–22.
8. Chen L, Jiang T, Li W, Geng S, Hussain S. Who should pay for online reviews? Design of an online user feedback mechanism. Electron Commer Res Appl. 2017;23:38–44.
9. Cambria E, Poria S, Gelbukh A, Thelwall M. Sentiment analysis is a big suitcase. IEEE Intell Syst. 2017;32(6):74–80.
10. Wang Z, Chong CS, Lan L, Yang Y, Ho S, Tong JC. Fine-grained sentiment analysis of social media with emotion sensing. Future Technol Conf. 2016:1361–4.
11. Xing FZ, Pallucchini F, Cambria E. Cognitive-inspired domain adaptation of sentiment lexicons. Inf Process Manag. 2019;56(3): 554–64.
12. Cambria E, Poria S, Hazarika D, Kwok K. SenticNet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings. In: The thirty-second AAAI conference on artificial intelligence (AAAI-18); 2018. p. 1795–802.
13. Mondal A, Cambria E, Das D, Hussain A, Bandyopadhyay S. Relation extraction of medical concepts using categorization and sentiment analysis. Cognit Comput. 2018;10(4):670–85.
14. Lauren P, Qu G, Yang J, Watta P, Huang G, Lendasse A. Generating word embeddings from an extreme learning machine for sentiment analysis and sequence labeling tasks. Cognit Comput. 2018;10(4): 625–38.
15. Li Y, Pan Q, Yang T, Wang S, Tang J, Cambria E. Learning word representations for sentiment analysis. Cognit Comput. 2017;9(6): 843–51.
16. Kolchyna O, Souza TTP, Treleaven P, Aste T. Twitter sentiment analysis: lexicon method, machine learning method and their combination. arXiv preprint arXiv. 2015:32.
17. Zhang L, Ghosh R, Dekhil M, Hsu M, Liu B. Combining lexicon-based and learning-based methods for twitter sentiment analysis. Int J Electron Commun Soft Comput Sci Eng. 2015;89:1–8.
18. Cambria E, Olsher D, Kwok K. Sentic activation: a two-level affective common sense reasoning framework. Proc AAAI. 2012: 186–92.
19. Cambria E, Mazzocco T, Hussain A, Eckl C. Sentic medoids: organizing affective common sense knowledge in a multi-dimensional vector space. LNCS. 2011;6677:601–10.
20. Wang Z, Tong JC, Ho SB. Method and system of intelligent sentiment and emotion sensing with adaptive learning. In: Patent

cooperation treaty (PCT) international application no.PCT/SG2017/050172; 2017.

21. Dashtipour K, Poria S, Hussain A, Cambria E. Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cognit Comput. 2016;8(4):757–71.
22. Narayanan V, Arora I, Bhatia A. Fast and accurate sentiment classification using an enhanced Naive Bayes model. Int Conf Intell Data Eng Automated Learn. 2013:194–201.
23. Wang Z, Tong JC, Chin HC. Enhancing machine-learning methods for sentiment classification of web data. Asia Inf Retr Symp. 2014;8870:394–405.
24. Chang W, Wang J. Mine is yours? Using sentiment analysis to explore the degree of risk in the sharing economy. Electron Commer Res Appl. 2018;28:141–58.
25. Al-obeidat F, Spencer B, Kafeza E. The opinion management framework: identifying and addressing customer concerns extracted from online product reviews. Electron Commer Res Appl. 2018;27:52–64.
26. Malandri L, Xing FZ, Orsenigo C, Vercellis C, Cambria E. Public mood – driven asset allocation: the importance of financial sentiment in portfolio management. Cognit Comput. 2018;10(6):1167–76.
27. Cambria E, Hussain A, Havasi C, Eckl C. SenticSpace: visualizing opinions and sentiments in a multi-dimensional vector space. Knowl-Based Intell Inf Eng Syst. 2010:385–93.
28. Tang J, Alelyani S, Liu H. Feature selection for classification: a review. Data Classif Algorithms Appl. 2014:37.
29. Duric A, Song F. Feature selection for sentiment analysis based on content and syntax models. Decis Support Syst. 2012;53(4):704–11.
30. Wang S, Li D, Song X, Wei Y, Li H. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. Expert Syst Appl. 2011;38(7):8696–702.
31. Al-Radaideh QA, Al-Qudah GY. Application of rough set-based feature selection for Arabic sentiment analysis. Cognit Comput. 2017;9(4):436–45.
32. Prusa JD, Khoshgoftaar TM, Dittman DJ. Impact of feature selection techniques for tweet sentiment classification. Twenty-Eighth Int Flairs Conf. 2015:299–304.
33. Nigam K, Lafferty J, Mccallum A. Using maximum entropy for text classification. In: IJCAI-99 workshop on machine learning for information filtering; 1999. p. 61–7.
34. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
35. Huang G, Zhu Q, Siew C. "Extreme learning machine: a new learning scheme of feedforward neural networks," in Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, 2004, vol. 2, no. August 2004, pp. 985–990.
36. Li S, Xia R, Zong C, Huang C-R. "A framework of feature selection methods for text categorization," in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009, no August, pp. 692–700.
37. Boiy E, Moens M-F. A machine learning approach to sentiment analysis in multilingual Web texts. Inf Retr Boston. Sep. 2009;12(5):526–58.
38. "Twitter-sentiment-analyzer," Available from: https://github.com/ravikiranj/twitter-sentiment-analyzer/tree/master/data [Cited 4 Sep. 2013].
39. "Twitter-sentiment-analysis2," Available from: https://www.kaggle.com/c/twitter-sentiment-analysis2/data [Cited 2 Dec. 2017].
40. Liu X, Gao C, Li P. A comparative analysis of support vector machines and extreme learning machines. Neural Netw. 2012;33:58–66.
41. Gelman A, Goodrich B, Gabry J, Ali I. R-squared for Bayesian regression models. Am Stat. 2018:1–6.