Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2018, Article ID 1407817, 5 pages
https://doi.org/10.1155/2018/1407817

Research Article

On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis

Asriyanti Indah Pratiwi and Adiwijaya

Telkom University, Telekomunikasi Street No. 1, Bandung 40257, Indonesia

Correspondence should be addressed to Asriyanti Indah Pratiwi; asriyantiindahpratiwi@gmail.com

Received 10 July 2017; Revised 9 October 2017; Accepted 26 November 2017; Published 19 February 2018

Academic Editor: Rodolfo Zunino

Sentiment analysis in a movie review is the needs of today lifestyle. Unfortunately, enormous features make the sentiment of analysis slow and less sensitive. Finding the optimum feature selection and classification is still a challenge. In order to handle an enormous number of features and provide better sentiment classification, an information-based feature selection and classification are proposed. The proposed method reduces more than 90% unnecessary features while the proposed classification scheme achieves 96% accuracy of sentiment classification. From the experimental results, it can be concluded that the combination of proposed feature selection and classification achieves the best performance so far.

1. Introduction

One of the interesting challenges in text categorization is sentiment analysis, a study that analyzes the subjective information of specific object [1]. Sentiment analysis can be applied on various level: document level, sentence level, and feature level.

Sentiment-based categorization in the movie review is a document-level sentiment analysis. It treats the review as a set of independent words by ignoring the sequence of words on a text. Every single unique word and phrase can be used as the document features. As a result, it constructs massive numbers of features. In addition, it also slows down the process and makes the classification task bias [2].

Actually, not all features are necessary. Most of the features are irrelevant to the class label. On the other hand, a good feature for classification is the one that has maximum relevance with the output class.

As feature selection in sentiment analysis is a crucial part, in this paper, we proposed an information gain based feature selection. In addition, we also proposed classification schemes based on the dictionary that is constructed by selected features.

2. Previous Work

There are two common approaches to sentiment analysis: machine learning methods and knowledge-based methods. Cambria [3] suggested the combination of both methods: using machine learning to provide the limitations of the sentiment knowledge. On the other hand, it cannot be applied in movie review. The sentiment knowledge such as SenticNet is highly dependent on domain and context. For example, "funny" means positive for comedy but negative for horror movie [4].

Machine learning-based sentiment analysis on movie review initialized by Pang et al. [5]. Their work performed 70%–80% accuracy while the human baselines sentiment analysis only reaches 70% accuracy. In 2014, Dos Santos and Gatti [6] used deep learning method for sentence-level sentiment analysis that reached 70%–85% accuracy. Words and characters are used as sentiment features. Unfortunately, the massive constructed features resulted in a long-time computation.

In order to provide robust machine learning classification, a feature selection technique is required [7]. Some researchers focus on reducing the number of features [8]. Manurung [9] proposed a feature selection scheme named

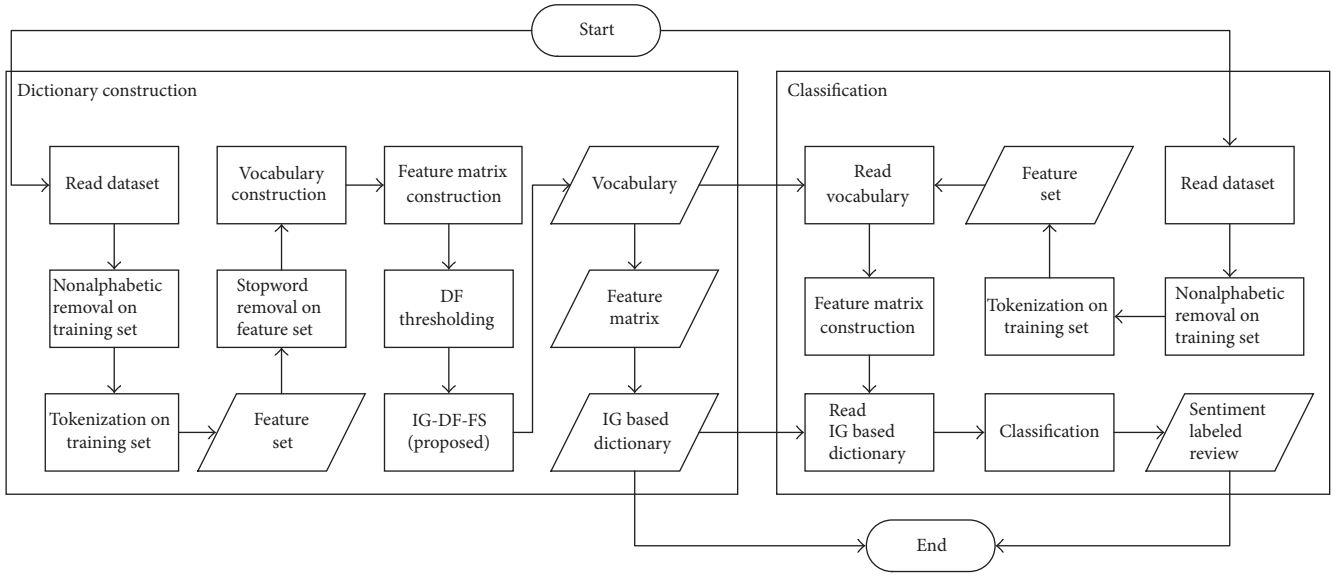

Figure 1: Classification flowchart.

feature-count (FC). FC selects n-top subfeatures with the highest frequency count. It only costs $O(n)$ to select the subfeatures. O then contrary, it may select a feature which has no relevance to the output class, since high occurrence does not indicate high relevance to the output class.

Nicholls and Song [8] research and OKeefe and Koprinska [10] research proposed similar idea to select features based on the difference between document frequency (DF) in class positive and DF in class negative. It was named Document Frequency Difference (DFD). DFD selects the feature that has the highest proportion between the positive DF-negative DF difference and the total number of documents. Their research may select feature which has high difference but less relevant to the output class.

Information theory-based feature selection such as information gain or mutual information was also proposed in sentiment analysis [11, 12]. In advance, Abbasi et al. proposed a heuristic search procedure to search optimum subfeature based on its information gain (IG) value named Entropy Weighted Genetic Algorithm (EWGA) [13]. EWGA search optimal subfeatures using Genetic Algorithm (GA) which its initial population is selected by information gain (IG) thresholding schemes. Compared to the other, EWGA is the most powerful feature selection so far. It selected features that achieved 88% accuracy of classification. However, it took high-cost computation.

This study uses polarity v.2.0 from Cornell review datasets, a benchmark dataset for document-level sentiment analysis, that consists of 1000 positive and 1000 negative processed reviews [14]. This dataset split into tenfold cross-validation.

3. Information Gain on Movie Review

Information gain measures how mixed up the features are [15]. In sentiment analysis domain, information gain is used to measure the relevance of attribute A in class C. The higher the value of mutual information between classes C and attribute A, the higher the relevance between classes C and attribute A.

$$I(C, A) = H(C) - H(C \mid A), \tag{1}$$

where $H(C) = -\sum_{cEC} p(C) \log p(C)$, the entropy of the class, and $H(C \mid A)$ is the conditional entropy of class given attribute, $H(C \mid A) = -\sum_{cEC} p(C \mid A) \log p(C \mid A)$. Since Cornell movie review dataset has balanced class, the probability of class C for both positive and negative is equal to 0.5. As a result, the entropy of classes $H(C)$ is equal to 1. Then the information gain can be formulated as

$$I(C, A) = 1 - H(C \mid A). \tag{2}$$

The minimum value of $I(C, A)$ occurs if only if $H(C \mid A) = 1$ which means attribute A and classes C are not related at all. On the contrary, we tend to choose attribute A that mostly appears in one class C either positive or negative. On the other words, the best features are the set of attributes that only appear in one class. It means the maximum $I(C \mid A)$ is reached when $P(A)$ is equal to $P(A \mid C_1)$ resulting in $P(C_1 \mid A)$ and $H(C_1 \mid A)$ being equal to 0.5. When $P(A) = P(A \mid C_1)$, then the value of $P(A \mid C_2)$ results in $P(C_2 \mid A) = 0$ and $H(C_1 \mid A) = 0$. The value of $I(C, A)$ is varied from 0 to 0.5.

4. Sentiment Analysis Framework

This study uses polarity v.2.0 from Cornell review datasets, a benchmark dataset for document-level sentiment analysis, that consists of 1000 positive and 1000 negative processed reviews [14]. This dataset split into tenfold cross-validation.

Figure 1 shows the process of proposed sentiment analysis. The process was categorized into dictionary construction phase and classification phase. Dictionary construction phase constructs a dictionary that can be used to classify the

```
(1) procedure IGDF–Feature–Selection(input: {array of attributes A and its class C},
    output: {positive and negative feature set})
(2)   for each features in featureset do
(3)      calculate I(C | A)
(4)   end for
(5)   for each IGscore in I(C | A) do
(6)      if I(C | A) == 0.5 then
(7)         Vocabulary ← Vocabulary + A
(8)         if P(A) == P(A | C_positive) then
(9)            featureset_positive ← featureset_positive + A
(10)        else
(11)           featureset_negative ← featureset_negative + A
(12)        end if
(13)     end if
(14)  end for
(15) end procedure
```

Algorithm 1: IGDF feature selection.

review: positive or negative. Here are the steps of dictionary construction phase in this study: (1) reading the dataset, (2) nonalphabetic removal, (3) tokenization, (4) stopwords removal, (5) stemming (optional), (6) initial vocabulary construction, (7) initial feature matrix construction, (8) DF thresholding, (9) IG-DF-FS, and (10) dictionary construction.

Similar to the dictionary construction phase, classification phase also consists of preprocessing and feature construction. On the contrary, it uses the constructed dictionary instead of selecting feature and constructs another dictionary. The result of this phase is sentiment labeled movie review.

4.1. IG-DF Feature Selection. Previous work on information gain [16] selects feature that has high relevance with the output class. Those features commonly appear in positive class or negative class only. Unfortunately, it may appear only a few times since the sentiment can be expressed in a various way. As a result, overfitting occurs since those features do not appear.

On the other hand, DF thresholding [8, 12] selects feature that appears most in the training set. It may select feature that always appears in both classes. Those features are unnecessary since it cannot differentiate the class to which it belongs.

In this study, we propose a combination of information gain and DF thresholding feature selection, named IGDFFS. IGDFFS selects a feature that has IG score equal to 0.5. It means those features highly related to one class only. These schemes succeed in reducing about 90% of unnecessary features (Algorithm 1).

4.2. Classification. As it is known that entropy and information gain are commonly used in decision tree. The selected feature with the highest information gain determines the class of the review. Based on this intuition, we categorize our vocabulary into the positive feature and negative feature. A review will be classified into positive review if most of the features are positive and vice versa (Algorithm 2).

5. Results and Analysis

Figure 2 shows the performance previous feature selection (FFSA) [16] and proposed feature selection (IGDFFS). The results show that IGDFFS selects better features.

Proposed method selects feature that has high relevance to the output class and also has the highest occurrence. As a result, generated feature matrix has less zero value. On the contrary, the previous method may succeed in selecting high relevant features but probably takes rare features. The rare feature does not appear in another movie review document in training set and may not appear in the testing set. As a result, the generated feature matrix consists of a lot of zero value. A lot of documents which have not any features are hard to be classified.

One of the feature selection objectives is to avoid overfitting. Actually, in this case, common machine learning techniques may result in overfitting. The reason is the feature matrix in testing set consists of a lot of zero values more than the feature matrix in training set. Since the features affect machine learning model, then it is hard for machine learning to fit the model to the feature matrix in the testing set.

Figure 3 summarizes the performance of SVM, ANN, and IG classifier. Unfortunately, SVM and ANN suffer from overfitting problems. Their testing accuracy fails in achieving 70% accuracy. Different to ANN and SVM, IGC is quite stable in any condition. IGC succeed in avoiding overfitting problems. It can be concluded that IGC as proposed classifier performs better than the current classifier.

Information gain value tells how mixed a feature to the class is. IG value reaches the highest value (0.5 in this case) when the feature belongs to one class only. It means when the feature appears we make sure that the label must be positive or negative. In this case, the IG value of selected feature achieves the maximum value on average (0.5) so, it can be used for automatic classification. The specialty of proposed classification scheme is the independence from mathematical model. Since proposed classification method

```
(1) procedure IG-based–Classifier(input: {Sentiment Feature Vector: Vocabulary
    × Number of Document}, output: {Sentiment Label: positive or negative})
(2)    for each document in featurevector do
(3)       for each vocabinVocabulary do
(4)          if vocab is positive – features then
(5)             positive ← positive + 1
(6)          else
(7)             negative ← negative + 1
(8)          end if
(9)       end for
(10)      if positive > negative then
(11)         class_label ← class_label + 'positive'
(12)      else
(13)         class_label ← class_label + 'negative'
(14)      end if
(15)   end for
(16) end procedure
```

Algorithm 2: IG-based classification.

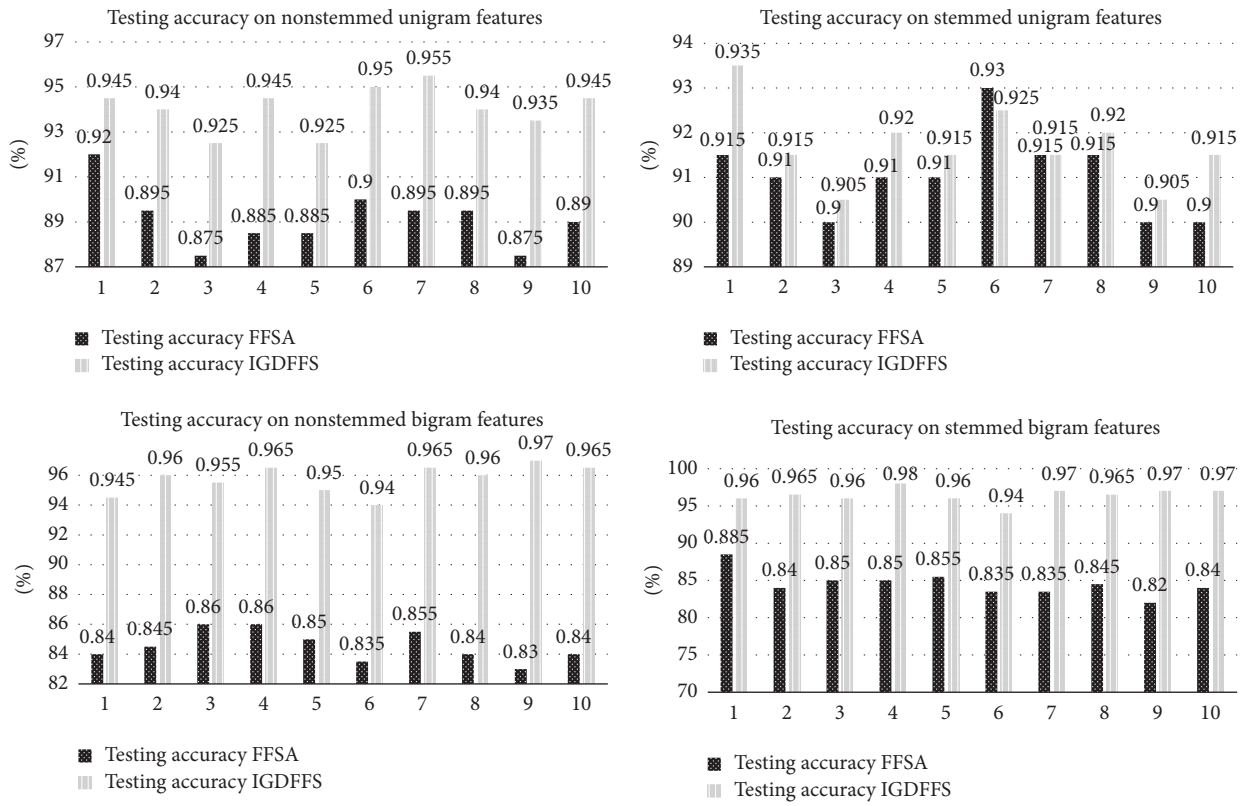

Figure 2: Feature selection performance comparison.

succeeds in avoiding overfitting, we can say that our method is better than the previous work.

6. Conclusion and Future Work

In order to provide better sentiment analysis system, an improvement of information gain based feature selection and classification was proposed. The proposed feature selection selects feature that has high information gain and high occurrence. As a result, it succeeded in providing feature that most probably appears in testing also. Proposed classifier used the positive and negative features obtained from the IG calculation before. Then, it takes less time than the previous classifier (SVM, ANN, etc.).

The combination of information gain and document frequency in this study proposed feature selection; IGDFFS selects subfeatures that satisfy these criteria: (1) high relevance to the output class and (2) high occurrence in

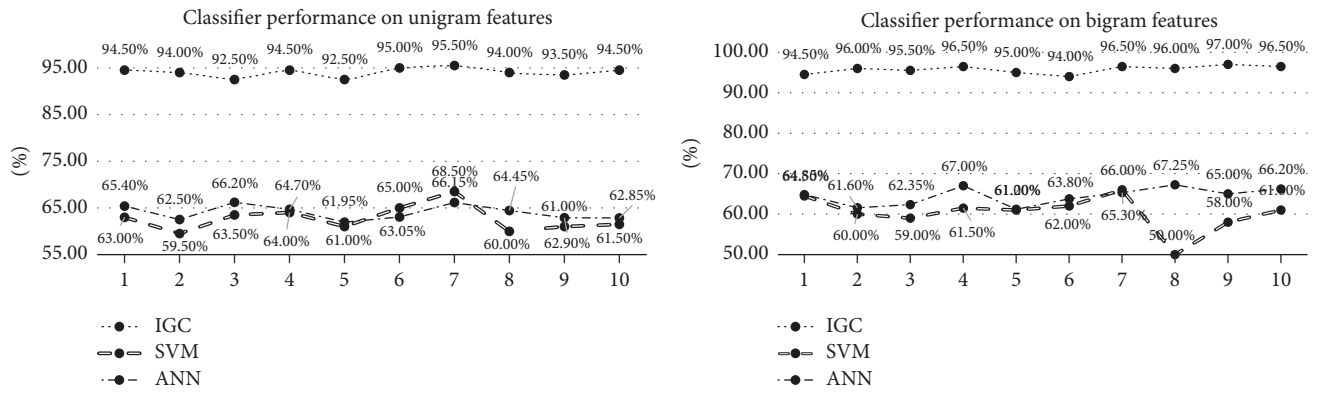

Figure 3: Sentiment classifier performance comparison.

dataset. As a result, it constructs subfeatures that reach better performance in the classification.

Compared to the current classifier, Information Gain Classifier (IGC) overcomes the recent high accuracy which belongs to EWGA (only 88.05%). It succeeded in avoiding overfitting problems in any condition. The performance of IGC is quite stable in both training and testing.

We are considering to groups the words based on their relevance to positive and negative reviews. Note that there are 171,476 words that are currently used and 47,156 obsolete words in English domain (based on Oxford English Dictionary). At least a finite number of groups would be less than the total number of words.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] B. Agarwal and N. Mittal, *Prominent Feature Extraction for Sentiment Analysis*, Springer, 2015.

[2] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 5, no. 4, pp. 537–550, 1994.

[3] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[4] P. Chaovalit and L. Zhou, "Movie review mining: a comparison between supervised and unsupervised classification approaches," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*, 112c pages, IEEE, 2005.

[5] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, pp. 79–86, Association for Computational Linguistics, Stroudsburg, Pa, USA, July 2002.

[6] C. N. Dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of the 25th International Conference on Computational Linguistics (COLING '14)*, pp. 69–78, 2014.

[7] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *FeaturE Extraction: Foundations and Applications*, vol. 207, Springer, 2008.

[8] C. Nicholls and F. Song, "Comparison of feature selection methods for sentiment analysis," in *Proceedings of the Canadian Conference on Artificial Intelligence*, pp. 286–289, Springer, 2010.

[9] R. Manurung, "Machine learning-based sentiment analysis of automatic indonesian translations of english movie reviews," in *Proceedings of the International Conference on Advanced Computational Intelligence and Its Applications (ICACIA)*, Depok, Indonesia, 2008.

[10] T. OKeefe and I. Koprinska, "Feature selection and weighting methods in sentiment analysis," in *Proceedings of the 14th Australasian document computing symposium*, pp. 67–74, Citeseer, Sydney, Australia, 2009.

[11] B. Agarwal and N. Mittal, "Text classification using machine learning methods-A survey," in *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012)*, vol. 236 of *Advances in Intelligent Systems and Computing*, pp. 701–709, Springer, India, December 2012.

[12] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques," *WSEAS Transactions on Computers*, vol. 4, no. 8, pp. 966–974, 2005.

[13] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: feature selection for opinion classification in Web forums," *ACM Transactions on Information and System Security*, vol. 26, no. 3, article 12, 2008.

[14] B. Pang and L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 271 pages, Association for Computational Linguistics, Barcelona, Spain, July 2004.

[15] R. M. Gray, *Entropy and Information Theory*, Springer Science and Business Media, 2011.

[16] F. Amiri, M. M. R. Yousefi, and C. Lucas, "Mutual information-based feature selection for intrusion detection systems," *Journal of Network & Computer Applications*, vol. 34, no. 4, pp. 1184–1199, 2011.