Article

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

Firstname Lastname 1, Firstname Lastname 2 and Firstname Lastname 2,\*

1 Affiliation 1; e-mail@e-mail.com

2 Affiliation 2; e-mail@e-mail.com

\* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials)

**Abstract:** Sentiment analysis for movie reviews is an increasingly important tool. Unfortunately, the massive number of features involved often causes sentiment analysis to become slow and less sensitive. Optimal feature selection and classification is still a significant challenge. To handle a large number of features and provide better sentiment classification, an information-based feature selection and classification method is proposed. The proposed method reduces unnecessary features by more than 90%, whereas the proposed classification scheme achieves 96% accuracy for sentiment classification. From the experimental results, it can be concluded that the proposed combined feature selection and classification method achieves better performance than other methods of this type.

|  |
| --- |
| **Citation:** Lastname, F.; Lastname, F.; Lastname, F. Title. *Appl. Sci.* **2021**, *11*, x. https://doi.org/10.3390/xxxxx  Academic Editor: Firstname Lastname  Received: date  Accepted: date  Published: date  **Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.    **Copyright:** © 2021 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). |

**Keywords:** keyword 1; keyword 2; keyword 3 (List three to ten pertinent keywords specific to the article yet reasonably common within the subject discipline.)

1. Introduction

One notable challenge in text categorization is sentiment analysis, a process that analyzes the subjective information of specific objects [1]. Sentiment analysis can be applied at various levels, that is, the document level, sentence level, and feature level.

Sentiment-based categorization in movie reviews involves document-level sentiment analysis. This method treats a review as a set of independent words by ignoring the sequence of words in the text. Every unique word and phrase can be used as a document feature. As a result, this type of sentiment analysis constructs a massive number of features. This abundance of features slows down the process and can introduce bias in the classification task [2].

However, not all features are necessary; most features are irrelevant to the class label. Thus, a good feature for classification is one that has high relevance to the output class.

As feature selection is a crucial component of sentiment analysis, in this paper, we propose an information gain (IG)-based feature selection method. In addition, we propose classification schemes based on the dictionary that is constructed by the selected features.

There are two common approaches to sentiment analysis: machine learning methods and knowledge-based methods. Cambria [3] suggested a combination of both methods, using machine learning to provide the limitations of sentiment knowledge. However, this technique cannot be applied to movie reviews. Sentiment knowledge, such as that provided by SenticNet, is highly dependent on domain and context. For example, the word “funny” has a positive connotation for a comedy movie, but a negative connotation for a horror movie [4].

Machine learning-based sentiment analysis of movie reviews was first performed by Pang et al. [5]. Their work achieved 70–80% accuracy, while the human baseline sentiment analysis method only reached 70% accuracy. In 2014, Dos Santos and Gatti [6] used a deep learning method for sentence-level sentiment analysis, reaching 70–85% accuracy. Words and characters were used as sentiment features. Unfortunately, the massive number of constructed features resulted in a long computation time.

To provide robust machine learning classification, a feature selection technique is required [7]. Some researchers have focused on reducing the number of features [8]. Manurung [9] proposed a feature selection scheme named feature-count (FC). FC selects the 𝑛-top subfeatures with the highest frequency count, an operation which has a time complexity of 𝑂(𝑛). However, this method may select a feature that has no relevance to the output class, since a high frequency of occurrence does not necessarily indicate high relevance to the output class.

The works of Nicholls and Song [8] and OKeefe and Koprinska [10] proposed a similar idea to select features based on the difference between document frequency (DF) in class positive and DF in class negative. This method was named Document Frequency Difference (DFD). DFD selects the feature that has the highest proportion between the positive vs. negative DF difference and the total number of documents. This approach may select features that have high differences in DF but are less relevant to the output class.

Information theory-based feature selection, using factors such as information gain or mutual information, has also been proposed for sentiment analysis [11, 12]. In advance, Abbasi et al. proposed a heuristic search procedure, named the entropy weighted genetic algorithm (EWGA), to search for optimal subfeatures based on their information gain values [13]. EWGA searches for optimal subfeatures using a genetic algorithm (GA) with an initial population selected using IG thresholding schemes. Compared to other options in this field, EWGA is the most powerful feature selection method to date. This approach selected features with 88% classification accuracy. However, it has a high computational cost.

This study uses the polarity dataset v2.0 from Cornell’s review datasets. This is a benchmark dataset for document-level sentiment analysis, consisting of 1000 positive and 1000 negative processed reviews [14]. This dataset was split for tenfold cross-validation.

2. Materials and Methods

2.1. Information Gain in Movie Reviews

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

𝐼 (𝐶, 𝐴) = 𝐻 (𝐶) − 𝐻 (𝐶 | 𝐴), (1)

where 𝐻(𝐶) = − ∑𝑐𝐸𝐶 𝑝(𝐶) log 𝑝(𝐶) is the entropy of the class and 𝐻(𝐶 | 𝐴) is the conditional entropy of the class given an attribute, 𝐻(𝐶 | 𝐴) = − ∑𝑐𝐸𝐶 𝑝(𝐶 | 𝐴) log 𝑝(𝐶 | 𝐴). Since the Cornell movie review dataset has balanced classes, the probability of class 𝐶 for both positive and negative results is equal to 0.5. As a result, the entropy of each class, 𝐻(𝐶), is equal to 1. Then, the information gain can be formulated as

𝐼 (𝐶, 𝐴) = 1 − 𝐻 (𝐶 | 𝐴). (2)

The minimum value of 𝐼(𝐶, 𝐴) occurs if and only if 𝐻(𝐶 | 𝐴) = 1, that is, attribute 𝐴 and class 𝐶 are not related at all. We tend to choose an attribute 𝐴 that mostly appears in one class 𝐶 as either positive or negative. For the other words, the best features are the set of attributes that only appear in one class. This means that the maximum 𝐼(𝐶 | 𝐴) is reached when 𝑃(𝐴) is equal to 𝑃(𝐴 | 𝐶1), resulting in 𝑃(𝐶1 | 𝐴) and 𝐻(𝐶1 | 𝐴) being equal to 0.5. When 𝑃(𝐴) = 𝑃(𝐴 | 𝐶1), then the value of 𝑃(𝐴 | 𝐶2) results in 𝑃(𝐶2 | 𝐴) = 0 and 𝐻(𝐶1 | 𝐴) = 0. The value of 𝐼(𝐶, 𝐴) varies from 0 to 0.5.

2.2. Sentiment Analysis Framework

This study uses the polarity dataset v2.0 from Cornell’s review datasets. This is a benchmark dataset for document-level sentiment analysis consisting of 1000 positive and 1000 negative reviews [14]. This dataset was split for tenfold cross-validation.

Figure 1 shows the proposed sentiment analysis process. The process was categorized into a dictionary construction phase and a classification phase. The dictionary construction phase constructs a dictionary that can be used to classify the review as positive or negative. The steps of the dictionary construction phase in this study are as follows: (1) reading the dataset, (2) nonalphabetic removal, (3) tokenization, (4) stopword removal, (5) stemming (optional), (6) initial vocabulary construction, (7) initial feature matrix construction, (8) DF thresholding, (9) information gain and DF thresholding feature selection (IGDFFS), and (10) dictionary construction.

|  |
| --- |
| Dictionary construction  Classification  Start  End  Vocabulary  Feature  set  Feature  matrix    Nonalphabetic  removal on  training set  Tokenization on  training set  Stopword  removal on  feature set  Vocabulary  construction  Feature matrix  construction  DF  thresholding  IG-DF-FS  )  proposed  (    Nonalphabetic  removal on  training set  Tokenization on  training set  Feature  set  Feature matrix  construction  IG based  dictionary  Sentiment  labeled  review  Read dataset  Read dataset  Read  IG based  dictionary  Read  vocabulary  Classification  Figure 1: Classification flowchart. |

Similar to the dictionary construction phase, the classification phase consists of preprocessing and feature construction. In contrast to the dictionary construction phase, it uses the constructed dictionary instead of selecting features and constructs another dictionary. This phase yields sentiment labeling of movie reviews.

2.3. IGDF Feature Selection

Previous work on information gain [16] selected features having high relevance to the output class. These features commonly appear in positive classes only or in negative classes only. Unfortunately, such features may appear only a few times, as a sentiment can be expressed in various ways. As a result, overfitting occurs because those features do not appear frequently.

In contrast, DF thresholding [8, 12] selects the features that appear most frequently in the training set. However, this method may select features that always appear in both classes. Such features are unnecessary, as the method cannot determine the classes to which these features belong.

In this study, we propose a combination of information gain and DF thresholding feature selection named IGDFFS. IGDFFS selects features that have IG scores equal to 0.5, indicating features highly related to one class only. This scheme succeeds in removing approximately 90% of unnecessary features (Algorithm 1).

|  |  |
| --- | --- |
| |  | | --- | | 1. **procedure** IGDF–Feature–Selection(input: {array of attributes 𝐴 and its class 𝐶}, output: {positive and negative feature set}) 2. **for** *each feature in featureset* **do** (3) *calculate* 𝐼(𝐶 | 𝐴) 3. **end for** 4. **for** *each IGscore in* 𝐼(𝐶 | 𝐴) **do** 5. **if** 𝐼(𝐶 | 𝐴) == 0.5 **then** 6. *Vocabulary* ←*Vocabulary* + 𝐴 7. **if** 𝑃(𝐴) == 𝑃(𝐴 | 𝐶𝑝𝑜𝑠𝑖𝑡𝑖V𝑒) **then** 8. *featuresetpositive* ← *featuresetpositive* + 𝐴 9. **else** 10. *featuresetnegative* ← *featuresetnegative* + 𝐴 11. **end if** 12. **end if** 13. **end for** 14. **end procedure** |   Algorithm 1: Information gain-document frequency (IGDF) feature selection. |

2.4. Classification

Entropy and information gain are commonly used in decision trees. The selected features with the highest information gain determine the class of the review. Based on this intuition, we categorize our vocabulary into positive and negative features. A review is classified as a positive review if most features are positive and vice versa (Algorithm 2).

|  |
| --- |
| 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. 𝑐𝑙𝑎𝑠𝑠𝑙𝑎𝑏𝑒𝑙 ← 𝑐𝑙𝑎𝑠𝑠𝑙𝑎𝑏𝑒𝑙 + 󸀠*positive*󸀠 12. **else** 13. 𝑐𝑙𝑎𝑠𝑠𝑙𝑎𝑏𝑒𝑙 ← 𝑐𝑙𝑎𝑠𝑠𝑙𝑎𝑏𝑒𝑙 + 󸀠*negative*󸀠 14. **end if** 15. **end for** 16. **end procedure** |

Algorithm 2: Information gain (IG) -based classification.

3. Results

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





Figure 2: Feature selection performance comparison.

The proposed method selects features that have both high relevance to the output class and the highest occurrence. As a result, the generated feature matrix has less zero value. In contrast, the previous method may succeed in selecting highly relevant features, but the selected features are likely to be rare. A rare feature does not appear in another movie review document in the training set and may not appear in the testing set. As a result, the generated feature matrix includes many zero values. Many documents without features are difficult to classify.

One feature selection objective is to avoid overfitting, which often results from common machine learning techniques. This is because the feature matrix in the testing set has many more zero values than the feature matrix in the training set does. Because these features affect machine learning models, it is difficult for machine learning to fit the model to the feature matrix in the testing set.

Figure 3 summarizes the performance of the SVM, ANN, and IG classifiers. Unfortunately, SVM and ANN suffer from overfitting and thus fail to achieve 70% accuracy. Unlike ANN and SVM, the information gain classifier (IGC) is quite stable in all conditions. IGC succeeds in avoiding overfitting; it can be concluded that using the proposed IGC offers performance better than that of the current classifier.

Figure 3: Sentiment classifier performance comparison.

Information gain value tells how mixed a feature to the class is. IG reaches the highest value (0.5 in this case) when the feature belongs to one class only. This means that when the feature appears, the label must be positive or negative. In this case, the IG of selected features achieves the maximum value (0.5) on average; thus, it can be used for automatic classification. The uniqueness of the proposed classification scheme lies in its independence from mathematical models. Since the proposed classification method succeeds in avoiding overfitting, we conclude that our method is more effective than those of previous works.

4. Discussion

To provide a better sentiment analysis system, a method of information gain-based feature selection and classification was proposed. The proposed method selects features with high information gain and high occurrence. As a result, it succeeded in providing features that were most likely to appear in testing as well. The proposed classifier used the positive and negative features obtained from the IG calculation before, performing its task more quickly than previous classifiers can (SVM, ANN, etc.).

A combination of information gain and document frequency was proposed for feature selection in this study. IGDFFS selects subfeatures that satisfy the following criteria: (1) high relevance to the output class and (2) high occurrence in the dataset. Thus, it constructs subfeatures that yield better classification performance.

Compared to current classifiers, the IGC has surpassed the high accuracy of EWGA (only 88.05%). The IGC succeeded in avoiding overfitting problems in diverse conditions, yielding stable performance in both training and testing.

For future work, we are considering grouping 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 the English domain (according to the Oxford English Dictionary). A limited number of groups would at least represent a dataset smaller than the total set of words.

**Author Contributions:** For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used “Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.” Please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

**Funding:** Please add: “This research received no external funding” or “This research was funded by NAME OF FUNDER, grant number XXX” and “The APC was funded by XXX.” Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding. Any errors may affect your future funding.

**Data Availability Statement:** In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Please refer to suggested Data Availability Statements in section “MDPI Research Data Policies” at https://www.mdpi.com/ethics. You might choose to exclude this statement if the study did not report any data.

**Acknowledgments:** In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

**Conflicts of Interest:** The authors declare no conflict of interest.

# 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.Pangand L.Lee,“Asentimentaleducation: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 informationbased feature selection for intrusion detection systems,” Journal of Network & Computer Applications, vol. 34, no. 4, pp. 1184–1199, 2011.