Scientific Journal of Informatics
Vol. 8, No. 2, November 2021
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Implementation of Support Vector Machine Algorithm with Correlation-Based Feature Selection and Term Frequency Inverse Document Frequency for Sentiment Analysis Review Hotel

Novia Puji Ririanti[1*], Aji Purwinarko[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Abstract.
Purpose: The study aims to reduce the number of irrelevant features in sentiment analysis with large features.
Methods/Study design/approach: The Support Vector Machine (SVM) algorithm is used to classify hotel review sentiment analysis because it has advantages in processing large datasets. Term Frequency-Inverse Document Frequency (TF-IDF) is used to give weight values to features in the dataset.
Result/Findings: This study's results indicate that the accuracy of the SVM method with TF-IDF produces an accuracy of 93.14%, and the SVM method in the classification of hotel reviews by implementing TF-IDF and CFS has increased by 1.18% from 93.14% to 94.32%.
Novelty/Originality/Value: Use of Correlation-Based Feature Section (CFS) for the feature selection process, which reduces the number of irrelevant features by ranking the feature subset based on the strong correlation value in each feature.

Keywords: Support Vector Machine, Correlation-Based Feature Section, Term Frequency-Inverse Document Frequency, Hotel Reviews

Received April 2021 / **Revised** June 2021 / **Accepted** October 2021

INTRODUCTION

The development of the Internet in the developing tourism sector has resulted in extensive reviews of hospitality and tourism services widely available on websites and other social networks. Reviews in the form of information are beneficial for people going on a tour or hotel booking. Some people tend to seek information from other visitors before deciding to make the right choice and make a hotel booking. Many platforms and websites are used as a facility for online opinions and experiences by internet users. Public opinion can be grouped or classified using sentiment analysis techniques. Sentiment analysis is a process that aims to determine the contents of a dataset in the form of text (documents, sentences, paragraphs) to be positive, negative, or neutral [1].

Sentiment analysis can also overcome text classification by automatically grouping user reviews into positive or negative opinions [2]. Opinions depicted are subjective expressions in the form of textual information that can describe a person's sentiments, opinions, or feelings about an event and nature. Textual information can be processed using a text mining process [3]. Text mining is also referred to as text mining, which is defined as the process of obtaining information from a set of text data [4]. The text mining process is divided into four categories: classification, analysis association, information extraction, and grouping.

Several classification algorithms are applied in sentiment analysis, such as Naïve Bayes (NB), Support Vector Machine (SVM), KNN. Several studies have been done to classify sentiment analysis against

*Corresponding author.
Email addresses: noviapuji04@students.unnes.ac.id (Ririanti), aji.purwinarko@mail.unnes.ac.id (Purwinarko)
DOI: 10.15294/sji.v8i2.29992

available reviews, and the SVM algorithm has become a famous classification and regression method for linear and nonlinear problems [5].

Before entering the classification stage, the words contained in the dataset are weighted using the Term Frequency Inverse Document Frequency (TF-IDF) technique. TF-IDF combines frequency term and document frequency-inverse to produce weights for each feature in the document [1]. With the many features possessed by data, sometimes some features may have irrelevant value for mining tasks, and if they include irrelevant features, it can harm and confuse the task of the classification algorithm [6].

Therefore, feature selection is necessary, which identifies and eliminates features with irrelevant or redundant values. The correlation-based feature selection (CFS) method is a simple filter algorithm and feature selection algorithm that ranks feature subsets and finds the benefits of features or feature subsets based on correlation [7]. The CFS will select the best feature subset containing features that strongly correlate with the target class but are not correlated with each other. However, if the data sample is limited, the attributes chosen by the CFS are not necessarily the attributes that give the best accuracy results [8]. This study uses a collection of hotel review datasets taken from https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe. The purpose of this study was to determine the accuracy of the SVM algorithm results after implementing CFS as feature selection and TFIDF as feature weighting for hotel review analysis sentiment.

METHODS

The steps in this study went through several stages, namely, text pre-processing, application of the TF-IDF feature as word weighting, application of CFS as feature selection, and the classification process using the SVM method. The study began by entering the Hotel Reviews dataset. The data then processed through the text pre-processing stage with tokenization, transform case, stopword removal, and stemming. The weighting of each word in the dataset was carried out using the TF-IDF technique. After the word weighting was carried out, the feature selection stage was conducted with the CFS, which functions to reduce irrelevant and low-value features to reduce the accuracy value in the classification algorithm. Based on the selected features, the classification process was carried out using the SVM algorithm. Then the classification model was tested using test data and evaluated using a confusion matrix to produce a value. The research method flow chart can be seen in Figure 1.

Dataset

The data used in this study were 2000 hotel review datasets with positive review labels and negative reviews, which can be downloaded from https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe.

Pre-processing

The text pre-processing stages carried out in this study were tokenization, transform Case, stopword removal, stemming. Text pre-processing is the stage of the initial process of the text to prepare the text into data that will be processed further.

1) Tokenization is the process of separating a set of words into pieces of words, phrases, symbols, or elements that have other meanings or are called tokens or terms [9].
2) Transform case stage can help normalize text. Text will be converted into a single letter form in uppercase or lowercase letters [10].
3) Stopword Removal, often referred to as a hyphen, is defined as a word that appears very often in text documents and does not give significant meaning to the document's contents [11].
4) The stemming stage is used to change the word into its basic form. The goal is to reduce the number of words that have primary forms to save time and memory space [12].

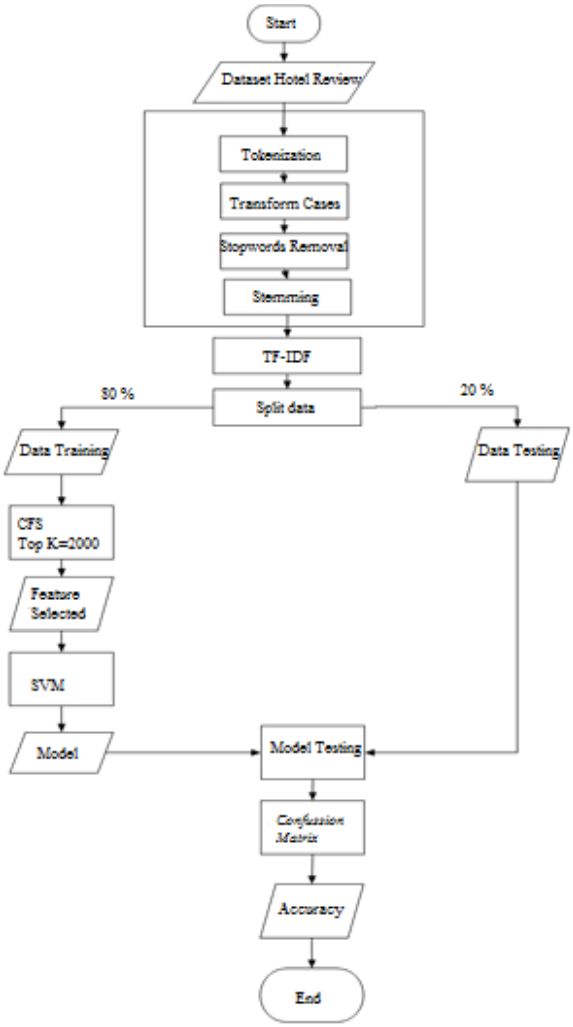

Figure 1. Support Vector Machine algorithm with TF-IDF and CFS flowchart

Term Frequency-Inverse Document Frequency (TF-IDF)

The TF-IDF method was used in the current study to determine the weight value of each word contained in each document [13]. TF-IDF is one of the most well-known algorithms used in text mining research. The first step in calculating TF-IDF was to count the appearance of words or Term Frequency (TF) in each document. Next, by calculating the Inverse Document Frequency (IDF) value. IDF value is the inverse of DF where the number of documents the word/term i appears. DF is the document appearance value of each word in each document. The IDF value is defined as equation 1.

$$IDF_t = \log \frac{N}{DF_t} \quad (1)$$

Where N is the total number of files in the document, and the formula for calculating TFIDF can be stated as equation 2.

$$TFIDF_t = TF_t \times IDF_t \quad (2)$$

Correlation-Based Feature Selection (CFS)

CFS is a filter algorithm and a simple feature selection algorithm that ranks feature subsets and finds the benefits of features or feature subsets based on correlation [7]. The CFS method will select the best feature subset containing features that strongly correlate with the target class but are not correlated with each other. However, if the data sample is limited, the attributes chosen by the CFS are not necessarily the attributes that give the best accuracy results [8]. The equation for calculating the correlation value can be formulated with equation 3.

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \quad (3)$$

CFS is also an automatic algorithm that does not require the user to specify the number of features to be selected [14].

Classification

Text classification was the process of classifying documents with predetermined text categories. It can be defined if di is a document from a document set and $\{C\_1, C\_2, C\_3 \ldots, C\_n\}$ is the set of all categories, then the text classification determines one category cj to the document. Based on its characteristics, documents can be labeled for one or more classes [15].

RESULT AND DISCUSSION

Results

This study used the python programming language. The classification process on the dataset used the SVM algorithm method, the classification in the hotel review dataset with the SVM algorithm, and uses the word weighting TF-IDF results in a value of 93.14%. The hotel review dataset classification with the SVM algorithm and using TF-IDF word weighting and the CFS selection feature produces an accuracy of 94.32%.

Pre-processing Result

The pre-processing stage serves to prepare the dataset so that it is ready to be processed into information. This pre-processing stage can provide optimal results in the classification process that will be used. The pre-processing stages used in this study are tokenization, transform case, stopword removal, and stemming. After pre-processing, the result is shown in Table 1.

Table 1. Pre-processing result

Review	Pre-processing Result
The floor in my room was filfy dirty Very basic rooms I had a 20yr old tv in my room Fridge did not work Overpriced breakfast	'floor' 'room' 'filfi' 'dirti' 'basic' 'room' 'old' 'tv' 'room' 'fridg' 'work' 'overpr' 'breakfast'
Comfy bed good location	'comfi' 'bed' 'good' 'locat'

TF-IDF Result

TF-IDF method is used to determine the weight value of each word contained in each document. The first step in calculating TF-IDF is to count the appearance of words or Term Frequency (TF) in each document. Next, by calculating the Inverse Document Frequency (IDF) value. The result of TF-IDF calculation is represented in Table 2.

Table 2. Result of TF-IDF

Feature	Weight
Floor	0.27380
Room	0.27380
Filfi	0.27380
Dirti	0.27380
Basic	0.27380
Old	0.27380
Tv	0.27380
Fridg	0.27380
Work	0.27380
Overpr	0.27380
breakfast	0.27380
Comfi	0.52335
Bed	0.52335
Good	0.52335
Locat	0.52335

CFS Result

After the TF-IDF process is carried out, a numeric feature is generated, which will then be carried out by the feature selection process using the CFS. The CFS method will select the features with the highest correlation weight value with the Top K value = 2000. The sample features that have the highest CFS weight can be seen in Table 3.

Table 3. Sample features that have a high CFS weight

feature	CFS
Polit	0.917961
Staff	0.917961
Style	0.921529
Quit	0.923860
Bed	0.921741
Warm	0.939528

Discussion

In this research, there will be two stages of the classification process. The first stage is to carry out the sentiment classification process on the hotel review dataset, which has gone through text pre-processing and TF-IDF with the SVM algorithm. The second is by applying CFS to the SVM algorithm.

Classification SVM+TF-IDF

The first application of classification is to use the SVM algorithm for classification and TF-IDF as feature weighting. The SVM algorithm is applied using a hotel review dataset with 2000 data gone through text pre-processing. The SVM and TF-IDF algorithms obtain an accuracy of 93.14%. With this accuracy, it is assessed that the SVM algorithm can classify the hotel review dataset well because it obtains an accuracy result more excellent than 60% or greater than the classification error rate. However, the results of this accuracy can still be improved by implementing CFS on SVM.

Classification SVM+TF-IDF+CFS

This classification applies a combination of TF-IDF as feature weighting and CFS as feature selection in the SVM classification algorithm for sentiment analysis. The features that have known their weight with TF-IDF will then look for a good correlation value that has been selected by the CFS algorithm, which will later affect the accuracy results of the classification process. The selection of 2000 features with the highest correlation value is based on experiments carried out in the classification using the SVM algorithm. By using 2000 features with the highest CFS ranking, the highest accuracy results are obtained. So it can be concluded that 2000 features have the most robust correlation weight value among the correlations with other features. The accuracy results obtained from this classification can be seen in Table 4.

Table 4. Accuracy results of the SVM with TF-IDF and CFS algorithms

CFS	Results of the SVM Algorithm + TF-IDF + CFS Accuracy
100	91.51%
500	92.81%
1000	93.87%
1500	93.74%
2000	**94.32%**
2500	93.83%

In this classification process, testing was carried out 10 times and obtained the results as shown in Table 5.

Table 5. Average results of the SVM classification experiment

Fold	SVM + TF-IDF			SVM + TF-IDF + CFS		
	True Predicted	False Predicted	Accuracy	True Predicted	False Predicted	Accuracy
1	310	21	93.66%	313	18	94.56%
2	296	35	89.43%	304	27	91.84%
3	312	19	94.26%	313	18	95.17%
4	311	20	93.96%	312	19	94.86%
5	306	25	92.45%	310	21	93.05%
6	312	19	94.26%	311	20	96.68%
7	316	15	95.47%	316	15	96.68%
8	313	18	94.56%	312	19	95.47%
9	298	33	90.03%	302	29	92.15%
10	308	22	93.33%	309	21	93.33%
	Average		93.14%	Average		94.32%

The results of the accuracy of the application of the SVM + TF-IDF + CFS algorithm classification model can increase the accuracy of the algorithm by 1.18%.

CONCLUSION

This paper has tested using the SVM method to classify hotel review sentiment using TF-IDF and CFS. The application of the CFS selection feature in this study eliminates some features that have irrelevant values by ranking the feature subset based on the solid value in each feature. The selection of features used in this classification is based on the feature ranking, with the acquisition criteria based on the better classification. The evaluation of the accuracy shows that the SVM algorithm with TF-IDF and the CFS selection feature results in a higher classification that outperforms the approach using the SVM algorithm with TF-IDF.

REFERENCES

[1] E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades, "Ontology-based sentiment analysis of Twitter posts," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4065–4074, 2013.

[2] Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, "Sentiment classification of Internet restaurant reviews written in Cantonese," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 7674–7682, 2011.

[3] V. Kotu & B. Deshpande, "Data Exploration. In Predictive Analytics and Data Mining," in *Predictive and Analysis*, 2015, Ch. 3, pp. 37-61.

[4] F. Zhang, H. Fleyeh, X. Wang, & M. Lu, "Construction site accident analysis using text mining and natural language processing techniques," *Autom. Constr.*, vol. 99, pp. 238–248, 2019.

[5] R. Moraes, J. F. Valiati, & W. P. Gavião Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Syst. Appl.*, vol. 40, pp. 621-633, 2013.

[6] B. Azhagusundari & A. S. Thanamani, "Feature Selection based on Information Gain," *Int. J. Innov. Technol. Explor. Eng.*, vol. 2, no. 2, pp. 2278-3075, 2013.

[7] P. Yildirim, "Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease," *Int. J. Mach. Learn. Comput.*, vol. 5, no. 4, pp. 258 –263 2015.

[8] I. Jain, V. K. Jain, & R. Jain, "Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification," *Appl. Soft Comput. J.*, vol. 62, pp. 203–215, 2018.

[9] C. D. Manning, P. Raghavan, & H. Schütze, "*Introduction to Information Retrieval*," Cambridge: Cambridge University Press, 2008.

[10] V. Kalra, & R. Aggarwal, "Importance of Text Data Preprocessing & Implementation in RapidMiner," in *Proc. First Int. Conf. Inf. Technol. Knowl. Manag.*, 2018, pp. 71 – 75.

[11] M. Shah, A. Monga, S. Patel, M. Shah, H. Bakshi, "A study of prevalence of primary dysmenorrhea in young students - A cross-sectional study," *Heal. J.*, vol. 4, no. 2, pp.1-4, 2013.

[12] S. Kannan, V. Gurusamy, S. Vijayarani, J. Ilamathi, M. Nithya, S. Kannan, & V. Gurusamy, "Preprocessing Techniques for Text Mining," *Int. J. Comput. Sci. Commun. Netw.*, vol. 5, no. 1, pp. 7–16, 2015.

[13] W. B. Trihanto, R. Arifudin, & M. A. Muslim, "Information Retrieval System for Determining The Title of Journal Trends in Indonesian Language Using TF-IDF and Naive Bayes Classifier," *Sci. J. Inform.*, vol. 4, no. 2, 179–190, 2017.
[14] S. Sasikala, S. Appavu, & S. Geetha, "Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set," *Appl. Comput. Inform.*, vol. 12, pp. 117–127, 2016.
[15] R. Jindal, R. Malhotra, & A. Jain, "Techniques for text classification: Literature review and current trends," *Weobology*, vol. 12, no. 2, pp. 1-28, 2015.