19. Improve the Accuracy of Support Vector Machine Using Chi Square Statistic and Term Frequency Inverse Document Frequency on Movie Review Sentiment Analysis

by Much Aziz Muslim

Submission date: 23-Jul-2019 02:56PM (UTC+0700)

Submission ID: 1154290546

File name: nverse_Document_Frequency_on_Movie_Review_Sentiment_Analysis.pdf (901.04K)

Word count: 4460

Character count: 23379

Scientific Journal of Informatics
Vol. 6, No. 1, May 2019

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Improve the Accuracy of Support Vector Machine Using Chi Square Statistic and Term Frequency Inverse Document Frequency on Movie Review Sentiment Analysis

Ukhti Ikhsani Larasati[1], Much Aziz Muslim[2], Riza Arifudin[3], Alamsyah[4]

[1,2,3,4]Department of Computer Science, FMIPA, Universitas Negeri Semarang
Email: [1]ukhtiikhsani010@gmail.com,[2]a212muslim@yahoo.co.id,[3]riza.arifudin@gmail.com,

Abstract

Data processing can be done with text mining techniques. To process large text data is required a machine to explore opinions, including positive or negative opinions. Sentiment analysis is a process that applies text mining methods. Sentiment analysis is a process that aims to determine the content of the dataset in the form of text is positive or negative. Support vector machine is one of the classification algorithms that can be used for sentiment analysis. However, support vector machine works less well on the large-sized data. In addition, in the text mining process there are constraints one is number of attributes used. With many attributes it will reduce the performance of the classifier so as to provide a low level of accuracy. The purpose of this research is to increase the support vector machine accuracy with implementation of feature selection and feature weighting. Feature selection will reduce a large number of irrelevant attributes. In this study the feature is selected based on the top value of K = 500. Once selected the relevant attributes are then performed feature weighting to calculate the weight of each attribute selected. The feature selection method used is chi square statistic and feature weighting using Term Frequency Inverse Document Frequency (TFIDF). Result of experiment using Matlab R2017b is integration of support vector machine with chi square statistic and TFIDF that uses 10 fold cross validation gives an increase of accuracy of 11.5% with the following explanation, the accuracy of the support vector machine without applying chi square statistic and TFIDF resulted in an accuracy of 68.7% and the accuracy of the support vector machine by applying chi square statistic and TFIDF resulted in an accuracy of 80.2%.

Keywords: SVM, Chi square statistic, TFIDF, Sentiment Analysis, Text Classification.

1. INTRODUCTION

Distribution of information supported by technological developments that better facilitate the public in obtaining information for free and in large numbers, one of which is textual information. Textual information can be categorized into two, namely the facts and opinions. Fact is an objective expression of an entity, event, or nature of an object. While opinion is a subjective expression that describes a person's sentiments, opinions, or feelings about an entity, event, and nature. Textual information can be processed using the text mining process.

The problems in data mining can be grouped into classification, regression, association analysis, anomaly detection, time series, and text mining [1]. Text mining is the application of data mining with input is text data can be documents, messages, e-mail or page of a website [1]. According to [2], text mining can be broadly defined as an intensive knowledge process where users interact with datasets using analytical tools. Text mining is also known as text data mining [3]. Text mining is similar to data mining, in fact a tool for data mining is designed for structured data from a database but text mining is designed for unstructured or semi-structured datasets such as word documents, emails, and more.

One of the problems associated with text mining is sentiments analysis. According to [4], sentiment analysis is a process that aims to determine the content of a dataset in the form of text (documents, sentences, paragraphs, etc.) to be either positive or negative. Sentiment analysis is usually implemented on three levels: sentence level, document level, and aspect level. The main purpose of the document level is to classify all documents or topics into positive or negative classes. Sentence levels are based on the polarity of each individual sentence [5]. More details are described in [6], the main purpose of the document level sentiment analysis is to classify the opinion of a document as a positive or a negative opinion based on several large documents with the same topic. Sentence level sentiment analysis, classifies sentiment in each sentence by identifying the sentence subjective or objectively. If the sentence is subjective, sentence level sentiment analysis will determine the sentence including a positive or negative opinion.

Public opinion becomes very important for industry players. In [7] mentions that sentiment analysis is used by industry players to know opinions about the industry's products in order to predict future sales. It is also mentioned in [8] that the movie's industry applies sentiment analysis to find out public opinion. According to [9] by looking at people's opinions, it can influence people's thinking on certain products so that people can deduce the quality of a particular product. The public can provide a review of a particular product through a website page. The review provided in the form of text of opinions, among others are review of cosmetic products, electronics, books [10] , food [11] , [12] and [10] movies, and so on. Movie production is one of the growing industries. One example of a site that provides a review of a movie product is the Internet Movie Database (IMDB). IMDB is a website page that deals with movie and movie production. IMDB provides complete information about the production of a movie, any cast in the movie, a brief synopsis of the movie, trailer link, release date, and reviews from other users. People use IMDB to know the quality of movies before buying or watching a movie, because other people's comments and movie ratings typically influence the level of interest in buying or watching the movie.

Data mining methods can be distinguished based on statistical approaches known as statistical methods and machine learning based on some techniques of supervised learning and unsupervised learning [13]. Some classification algorithms are applied in sentiment analysis such as, [11] using Naïve Bayes (NB) and Support Vector Machine (SVM) to classify restaurant review sentiments. In [14] using four machine learning methods, namely NB, ME, Stochastic Gradient Descent, and SVM. SVM is a widely used method in text classification [13]. SVM is a fast and effective method for text classification [2]. According to [15], one of the problems in text classification or text data processing is the number of features/attributes used on a dataset that will degrade the performance of the classifier. To optimize the work of the classifier needs to be done by selecting relevant features using feature selection. Feature selection is used to reduce feature/attribute dimension by removing irrelevant words so as to improve classification accuracy. On [16] explains that the feature selection method is used to reduce the dimension of the dataset by removing features/attributes that are irrelevant for classification. Feature selection provides several advantages such as smaller dataset sizes, less computing requirements for text classification algorithms. On [17] explains that feature selection can be divided into two types, namely filter and wrapper. Examples of filter types are chi square, information gain, and log like hood ratio. Examples of wrapper types are forward selection and backward elimination.

Chi square statistic gives good results when combined with SVM algorithm [18]. Chi square statistic was used to test the independence of two events. For feature selection two events are term (ti) and class (Ck) [19]. Research [19] using chi square as a feature selection in the support vector machine algorithm, provides an effective result in Arabic dataset classification with an F-measure of 88.11.

After the feature is selected then the process of weighting feature (feature weighting). Weighting feature is done to weigh the weight of each feature. One method of feature weighting is the Term Frequency Inverse Document Frequency (TFIDF). TFIDF is a combination of the term frequency and inverse document frequency to generate weights for each term in each document [20]. Research [21] apply TFIDF to calculate the connectedness weight of a term against a document.

The purpose of this study is to improve the accuracy of the SVM with mene r apkan chi square and TFIDF. Based on the description above, to reduce the number of large attributes need to apply feature selection to perform the process of choosing the right attributes and reduction of the number of attributes to improve accuracy. To improve the accuracy of existing models, it is proposed chi square statistic as feature selection and feature weighting with TFIDF. Feature selection and feature weighting will be integrated with the SVM classification method.

2. METHODS

This research was conducted in several stages, according to the process of classification of text according to Figure 1. Classification is one technique in data mining [22]. Data analysis is an effort to work with data, organize data, sort it into manageable units, synthesize it, search for and find what is important and what must be studied and decided. This study consists of several stages: stage preprocessing, feature selection, weighting of the feature, and classification of sentiment analysis.

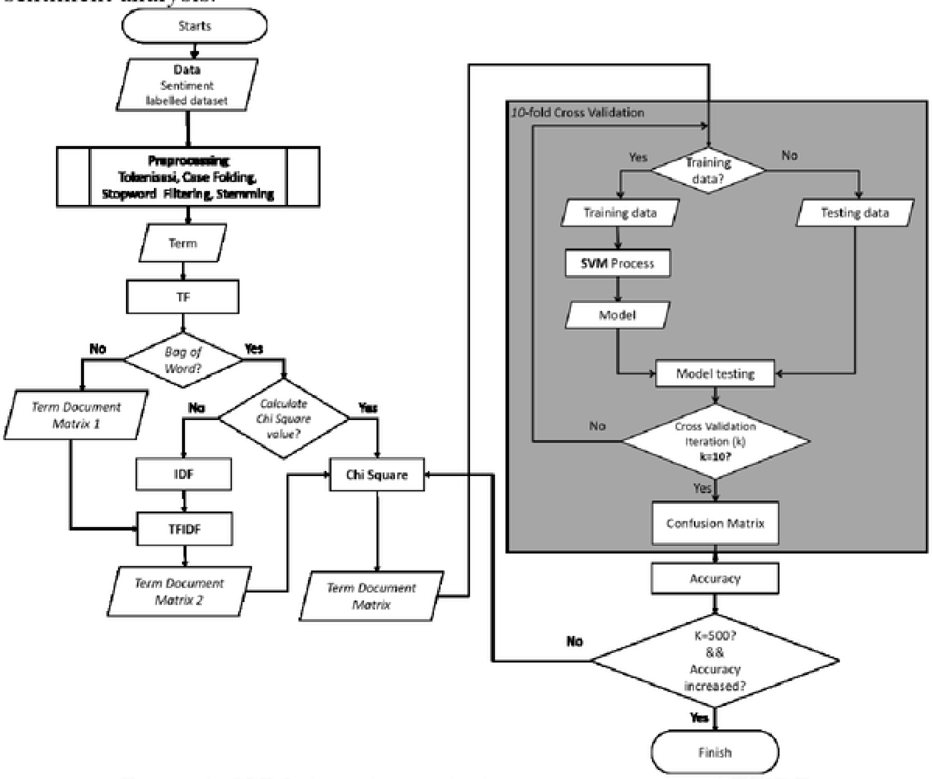

Figure 1. SVM algorithm with chi square statistic and TFIDF

2.1. Preprocessing

The preprocessing stage aims to prepare unstructured text documents ready for use for further processing. Preprocess stage conducted in this research is tokenize, case folding, stopword filtering, case folding, and stemming. Stopword filtering using stopword list is English stoplist and stemmer used is porter stemmer.

2.2. Feature Selection

The selection of attributes/features in this study using chi square statistic. Feature is selected based on the top value of K that is a number of K words with the highest chi square value. The top value of K is determined by the researcher, K = 500. Then it is repeated as much as K until the highest value of the classification is obtained. The calculation of chi square statistic for the selection of relevant attributes begins with a set of data that has been labeled class (positive, negative). Then from the data set with the concept of bag of word that contains the number of occurrence of term/word on each document for each class (positive, negative) as in Table 1.

Table 1. Bag-of- word (example)

Attribute name	*Total Occurrence*	*Document Occurrence*	pos	neg
good	42	41	37	5
bad	2	2	0	2
charact	2	2	1	1
bore	3	3	1	2
waste	2	2	0	2

1.2.1 *Chi Square Statistic*

Chi square statistic was used to test the independence of two events. For feature selection two events are term (ti) and class (Ck) [19]. "a" is the number of records/instance category Ck containing term ti, "b " is the number of records/instances which is not a category/class Ck containing term ti, "c" is the number of records/instances in the category Ck which contains no term ti, and "d" is the number of records/instances which is not a category Ck that does not contain the term ti. Where N is the entire document used. The chi square value for each term is calculated by Equation 1.

$$\chi^2(t_i, C_k) = \frac{N\times(ad-bc)^2}{(a+c)\times(b+d)\times(a+b)\times(c+d)} \qquad (1)$$

If $\chi^2(t_i, C_k) = 0$, term t_i and the category C_k independent; therefore term t_i has no effect on the category. The greater value of $\chi^2(t_i, C_k)$, then term t_i increasingly also affect the category. Selecting attribute/feature in this study using chi square statistics with the following steps.

1. Preparing bag-of-word results from the preprocessing stage as shown in Table 1 which displays some of the terms to be calculated for the chi square value.
2. Calculating the chi square value of each term in the bag of word obtained from the preprocessing stage using Eq. 1.
3. From Equation 1 this study uses the following values.
   - N = 1000, p = 500, dan n = 500; where p is the number of documents labeled positive and n is the number of documents labeled negative.

- C_k = data class (positive and negative).
- The value of "a" is the number of positive documents containing the term t_i in the pos column of Table 1.
- The value of b is the number of non-positive documents containing the term t_i in the neg column Table 1.
- The value of c is the number of positive documents but does not contain words/terms t_i that is $p - a$.
- The value of d is the number of documents that are not positive and do not contain the word/term t_i that is $n - b$.

4. Then the value of chi square for each term can be calculated by Equation 1 as follows. For example used a term that is "good".

$$\chi^2(t_i, c_k) = N\frac{(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

$$\chi^2(good, pos) = 1000\frac{((37x499)-(1x500))^2}{(0+500)(1+499)(0+1)(500+499)}$$

$$\chi^2(good, pos) = 1000\frac{(18315-2315)^2}{(500)(500)(42)(958)}$$

$$\chi^2(good, pos) = 25{,}45$$

Did t the value of *chi square* for the *term* "good " was 25.45.

5. Then each *term* in *bag-of-word is* sorted by *chi square* value from highest to lowest.
6. Feature selected based on the value of the top K is a K word of bag-of-word with the highest value of chi square. The study determines the value of K=500. Then the selected feature is the term with the highest 500 chi square value.

2.3. Feature Weighting

The weight of each feature for each document is calculated using TFIDF. The value of TFIDF is the weight of each feature on each document. After the preprocessing phase is completed and the relevant feature has been selected with feature selection there will be a number of N features that can be represented in the order of t1, t2, ..., tN. The ith document can be represented by a set of N-dimensional vector sequence is written to (Xi1, Xi2, ..., XiN) where Xij is the weight that calculates the level of interest term to j in the ith document. Vector space model is the result of the process of weighting each word in this case the word has become a feature that has been selected. One method of weighting is the TFIDF. This method calculates the Term Frequency (TF) and Inverse Document Frequency (IDF) values for each feature selected for N documents. TF value is defined by TF = tij, that is the number/total term i appears in document j. The DF value is the sum/total of the documents where the term i appears, this value is used to calculate the IDF. The IDF value is defined as in Equation 2.

$$IDF = log\left(\frac{N}{DF}\right) \quad (2)$$

where N is the number of the entire document. TFIDF is calculated by multiplying the Term Frequency (TF) with the Inverse Document Frequency (IDF) as in Equation 3.

$$TFIDF = TFx\,log\left(\frac{N}{DF}\right) \tag{3}$$

2.4. Mining Process

In the classification phase of sentiment analysis using SVM based on 10- fold cross validation training data and test data divided on 10- fold cross validation iteration [23], so that the learning and testing stage is done in 10-fold cross validation iteration as follows.

1) Prepare the dataset. The data used is the new *term document matrix* that is the result of the *feature* weighting stage.
2) Dividing data. Data is shared as much 10 equal parts. For the first iteration the test data used is as much as one piece of data and the other part as training data.
3) SVM process. At this stage training is done to get the classification model with training data based on the division in step 2). The modeling stages of SVM algorithm are as follows.
    a) Specifying the data point: $x_i = \{x_1, x_2, \ldots, x_n\} \in R^n$; R^n is a feature space with as many as *n* features.
    b) Specifying the class data: $y_i \in \{-1, +1\}$
    c) Pairing the data and class: $\{(x_i, y_i)\}_{i=1}^{N}$
    d) Minimize *margin* to determine *w* and *b* values

    $$\frac{1}{2}\|w\|^2 \text{ with } y_i\big((w.x_i) + b\big) \geq 1, i = 1, \ldots, l$$

    e) Specifies the separation hyperplane written as follows.
        (1) Inizialitation of $\alpha_i = 0,\ C = 1, gamma = 0.5, lamda = 0.5$
        Calculate the matrix $D_{ij} = y_i y_j\big(K(x_i, x_j) + \lambda^2\big)$
        (2) Perform step (a), (b), dan (c) belo for $i = 1,2, \ldots, l$
            (a) $E_i = \sum_{j=1}^{l} \alpha_j D_{ij}$
            (b) $\delta\alpha_i = min\{max[\gamma(1 - E_i), -\alpha_i], C - \alpha_i\}$
            (c) $\alpha_i = \alpha_i + \delta\alpha_i$

        Go back to step 2 to the value α reach convergent (no significant change).

4) Testing model with test data with decision function:

$$f(x) = sign((w.x) + b) \text{ or } f(x) = sign\left(\sum_{i=1}^{m} \alpha_i y_i K(x, x_i) + b\right)$$

5) Doing looping up to *k* looping in step 2) up to 4).
6) The final result will be obtained an output that is the level of accuracy. The accuracy level is obtained from the average on each iteration.

3. RESULT AND DISCUSSION

In this section, the experimental results are analyzed to evaluate the performance of the proposed data mining algorithm. The data used is movie review on Sentiment Labelled Sentences [24] taken from the UCI repository of machine learning. The first step is preprocessing data. In the preprocessing stage/preprocessing do case folding, tokenize, stopword filtering, and stemming. The preprocessing stage aims to prepare unstructured text documents to be ready for use in the next process. At this stage generated term document matrix and bag of word with as many as 2477 term candidate feature of the 1000 documents used and will be used at a later stage.

The next step is to choose relevant feature to the method of feature selection that is chi square statistic. By using Equation 1 the value of chi square is calculated for each candidate feature. The value of chi square is then sorted from the highest value to the lowest value. Chi square is applied to reduce a large number of attributes by taking a number of K=500 attribute of the highest ranking. Then we calculated the weight of each selected attribute with feature weighting TFIDF.

The next stage after the selected feature has been determined and has calculated the value of weight with TFIDF then the process of sentiment analysis with support vector machine algorithm can be done. At this stage will be calculated the highest level of support vector machine based on 10 fold cross validation in classification analysis of movie review sentiment with the application of chi square statistic and TFIDF. The level of accuracy of support vector machine algorithm on the classification of movie review analysis without chi square statistic and TFIDF treatment was 68.7%. After being given chi square treatment statistic and TFIDF support vector machine algorithm achieved the highest level of accuracy when the top value of K = 212 is 80.2% with an accuracy increase of 11.5%. This study used chi square statistic as feature selection based on the top value of K = 500. The update in this research is from the predetermined top K value, it is looping as much as K looping to get optimum accuracy value. The feature is selected based on the chi square value, the higher the chi square value the more relevant the feature. The preprocessing stage also plays a role in improving accuracy in this study. Most of the irrelevant features have been removed at the preprocessing stage. Table 1 shows the accuracy of the support vector machine in text classification for the movie review sentiment analysis without applying chi square and TFIDF and the accuracy of the support vector machine that implements chi square and TFIDF.

Table 1. Research result

Classifier	Accuracy
SVM	68,7%
SVM+Chi Square+TFIDF	80,2%

Application of chi square statistic and TFIDF on the support vector machine algorithm proved to be a good enough model to improve the accuracy of the support vector machine in the analysis of movie review sentiment in the data sentiment labelled dataset. With a given level of accuracy, this model is expected to be able to analyze the sentiment of the review data different with exact. So for further research, this model can be used for classification of sentiment analysis of other movie review data. This statement is aligned with [25], ie the level of accuracy (performance) increases with the application of feature selection. The study [25] applies a modification of the conventional IG feature selection that SAIG (Sparsity Adjusted Information Gain) shows improved classifier accuracy. Of the two datasets were used (amazon datasets and dataset movie (sentiment labeled dataset)) SAIG give better results than the IG on SVM and KNN. SQM + SAIG accuracy rate 67,9% (60 feature), SVM + IG 66,6% (100 feature), KNN + SAIG 68,2% (50 & 60 feature), and KNN + IG 57,4% (80 feature). However, with the same data this study shows better performance by applying conventional feature selection that is chi square statistic which shows an accuracy increase of 11.5% from SVM accuracy without application of chi square statistic and TFIDF 68,7% to 80,2% with the application of chi square statistic and TFIDF. Comparison of performance (accuracy) with previous research is presented in Table 2.

Table 2. Performance (accuracy) comparison

Author	*Classifier*	*Cross Validation*	*Feature Selecion*	*Baseline Accuracy*	*Best Accuracy*
On [26]	UPNN	-	No	40.5%	43.5%
On [25]	SVM	5	Yes	62.7%	67.9%
	NB	5	Yes	60.5%	60.0%
	KNN	5	Yes	57.4%	68.2%
On [27]	PFM	-	No	73.8%	78.9%
Proposed	SVM	10	Yes	68.7%	80.2%

4. CONCLUSION

This research uses Sentiment Labelled Data set taken from UCI Repository consists of 500 documents labeled positive and 500 documents labeled negatives. From result of experiment by using chi square with value of top K = 500 got highest accuracy value at top K = 212 and TFIDF, support vector machine showed an increase accuracy by 11, 5% from 68.7% to 80.2%. It can be concluded that the application of chi square statistic and TFIDF increases the accuracy of the support vector machine in the classification movie review sentiment analysis. Limitations in the sentiment analysis classification is highly dependent on the data to be tested. So this research can be used as a reference for further research by maximizing the data to be used in order to provide a more accurate level of accuracy. In this research use bag of word concept so that the feature/attributes formed are word by word, analysis is done based on word per word from given data and does not apply emoticon detection and negation detection. In addition to the data used, the need for further research is to find out how the results of this study (performance) can be used as supporting decision makers in movie production.

5. REFERENCES

[1] Kotu, V., & Deshpande, B. (2015). *Predictive Analytics and Data mining: Concepts and Practice with RapidMiner*. Waltham, MA: Elsevier/Morgan Kauffmann.

[2] Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

[3] Gupta, V., & Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, 1(1), 60-76.

[4] Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert System with Application*, 40, 4065-4074.

[5] Tripathy, A., Agrawal, A., & Rath, S. K. (2015). Classification of Sentimental Reviews Using Machine Learning Techniques. *Procedia Computer Science*, 57, 821-829.

[6] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Engineering Journal*, 5(4), 1093-1113.

[7] Liu, Y., Huang, X., An, A., & Yu, X. (2007). ARSA: A SentimentAware Model for Predicting Sales Performance Using Blogs. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*.

[8] Tsou, Benjamin K., Zhu, J., Wang, H., Zhu, M., & Ma, M. (2011). Aspect-Based Opinion Polling from Customer Reviews. *IEEE Transactions on Affective Computing*, 2(1), 37-49.

[9] Koh, N. S., Hu, N., & Clemons, E. K. (2010). Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications*, 9(5), 374–385.

[10] Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification an empirical comparison between SVM and ANN. *Expert System with Application*, 40, 621-633.

[11] Zhang, Z., Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment Classification of Internet Restaurant Reviews Written in Cantonese. *Expert Systems with Applications*, 38(6), 7674-7682.

[12] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs Up: Sentiment Classification Using Machine Learning Techniques. *EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing*.

[13] Jindal, R., Malhotra, R & Jain, A. (2015). Techniques for text classification: Literature review and current trends. *Weobology*, 12(2), 1-28.

[14] Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of Sentiment Reviews Using N-Gram Machine Learning Approach. *Expert Systems with Applications*, 57, 117-126.

[15] Wang, S., Li, D., Song, X., Wei, Y., & Li, H. (2011). A Feature Selection Method Based on Improved Fisher's Discriminant Ratio for Text Sentiment Classification. *Expert Systems with Applications*, 38(7), 8696-8702.

[16] Vala, M., & Gandhi, J. (2015). Survey of Text Classification Technique and Compare Classifier. *International Journal of Innovative Research in Computer and Communication Engineering*, 3(11), 10809-10813.

[17] Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X., & Wang, S. (2011). An Improved Particle Swarm Optimization for Feature Selection. *Journal of Bionic Engineering*, 8(2), 191-200.

[18] Meesad, P., Boonrawd, P., & Nuipian, V. (2011). A Chi-Square-Test for Word Importance Differentiation in Text Classification. *Proceedings of International Conference on Information and Electronics Engineering*.

[19] Mesleh, A. M. (2007). Chi Square Feature Extraction Based SVMs Arabic Language Text Categorization System. *Journal of Computer Science*, 3(6), 430-435.

[20] Manning, Christopher D., Prabhakar R., & Hinrich S. (2009). *An Introduction to Information Retrieval*. England: Cambridge University Press.

[21] Trihanto, W. B., R. Arifudin, & M. A. Muslim. (2017). Information Retrieval System for Determining The Title of Journal Trends in Indonesian Language Using TF-IDF and Naive Bayes Classifier. *Scientific Journal of Informatics*, 4(2), 180.

[22] Muslim, M. A., A. J. Herowati, E. Sugiharti, & B. Prasetiyo. (2018). Application of The Pessimistic Pruning to Increase The Accuracy of C4.5 Algorithm in Diagnosing Chronic Kidney Disease. *Journal of Physics: Conference Series 983* (1).

[23] Muslim, M. A., S. H. Rukmana, E. Sugiharti, B. Prasetiyo, & S. Alimah. (2018). Optimization of C4.5 Algorithm-based Particle Swarm Optimization for Breast Cancer Diagnosis. *Journal of Physics: Conference Series 983* (1).

[24] Kotzias, D., M. Denil, N. D. Freitas, & P. Smyth. (2015). From Group to Individual Labels using Deep Features. KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney: International Conference on Knowledge Discovery and Data Mining.

[25] Ong, B. Y., S. . Goh, & CC. Xu. (2015). Saprsity Adjusted Information Gain for Feature Selection in Sentiment Analysis. *Proceeding of IEEE International Conference on Big Data*. pp. 2122.

[26] Tang, D., B. Qin, & T. Liu. (2015). Learning semantic representations of users and products for document level sentiment classification. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. 1, 1014.

[27] Wang S., M. Zhou, G. Fei, Y. Chang, B. Liu. (2018). Contextual and Position-Aware Factorization Machines for Sentiment Classification. *arXiv preprint arXiv:1801.06172*.

19. Improve the Accuracy of Support Vector Machine Using Chi Square Statistic and Term Frequency Inverse Document Frequency on Movie Review Sentiment Analysis

ORIGINALITY REPORT

21%	%	21%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Jenie Sundari, Hamimah, Popon Handayani, Yunita et al. "Expert System To Detect Human’s Skin Diseases Using Forward Chaining Method Based On Web Mobile", MATEC Web of Conferences, 2018 Publication	2%
2	Meesala Shobha Rani, S. Sumathy. "Analysis on various machine learning based approaches with a perspective on the performance", 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), 2017 Publication	2%
3	Mihuandayani, Ema Utami, Emha Taufiq Luthfi. "Text mining based on tax comments as big data analysis using SVM and feature selection", 2018 International Conference on Information and Communications Technology (ICOIACT), 2018 Publication	1%
	Lecture Notes in Computer Science, 2014.	

4	Publication	1%
5	Studies in Computational Intelligence, 2016. Publication	1%
6	Rosy Indah Permatasari, M. Ali Fauzi, Putra Pandu Adikara, Eka Dewi Lukmana Sari. "Twitter Sentiment Analysis of Movie Reviews using Ensemble Features Based Naïve Bayes", 2018 International Conference on Sustainable Information Engineering and Technology (SIET), 2018 Publication	1%
7	Dinda Ayu Muthia, Dwi Andini Putri, Hilda Rachmi, Artika Surniandari. "Implementation of Text Mining in Predicting Consumer Interest on Digital Camera Products", 2018 6th International Conference on Cyber and IT Service Management (CITSM), 2018 Publication	1%
8	Lecture Notes in Computer Science, 2010. Publication	1%
9	Hui Zhang, Huguang Rao, Junzheng Feng. "Product innovation based on online review data mining: a case study of Huawei phones", Electronic Commerce Research, 2017 Publication	1%
10	Akshi Kumar, Vikrant Dabas, Parul Hooda. "Text classification algorithms for mining	1%

	unstructured data: a SWOT analysis", International Journal of Information Technology, 2018 Publication	
11	Studies in Computational Intelligence, 2014. Publication	1%
12	"Foundations of Intelligent Systems", Springer Science and Business Media LLC, 2014 Publication	1%
13	Muhammad Afzaal, Muhammad Usman, Alvis C.M. Fong, Simon Fong. "Multiaspect-based opinion classification model for tourist reviews", Expert Systems, 2019 Publication	1%
14	Zohaib Mushtaq, Akbari Yaqub, Ali Hassan, Shun Feng Su. "Performance Analysis of Supervised Classifiers Using PCA Based Techniques on Breast Cancer", 2019 International Conference on Engineering and Emerging Technologies (ICEET), 2019 Publication	<1%
15	Mahima Goyal, Vishal Bhatnagar. "chapter 10 A Classification Framework on Opinion Mining for Effective Recommendation Systems", IGI Global, 2017 Publication	<1%
16	Shoji Takimura, Ryosuke Harakawa, Takahiro Ogawa, Miki Haseyama. "Twitter Followee Recommendation Based on Multimodal FFM	<1%

Considering Social Relations", 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), 2018
Publication

17 "Pattern Recognition and Machine Intelligence", Springer Science and Business Media LLC, 2017
Publication
<1%

18 Tata Sutabri, Syopiansyah Jaya Putra, Muhammad Ridwan Effendi, Muhamad Nur Gunawan, Darmawan Napitupulu. "Sentiment Analysis for Popular e-traveling Sites in Indonesia using Naive Bayes", 2018 6th International Conference on Cyber and IT Service Management (CITSM), 2018
Publication
<1%

19 Shrawan Kumar Trivedi, Shubhamoy Dey. "Analysing user sentiment of Indian movie reviews", The Electronic Library, 2018
Publication
<1%

20 J. R. Méndez. "A Comparative Impact Study of Attribute Selection Techniques on Naïve Bayes Spam Filters", Lecture Notes in Computer Science, 2008
Publication
<1%

21 Priyanka Ingole, Smita Bhoir, A.V. Vidhate. "Hybrid Model for Text Classification", 2018 Second International Conference on Electronics, Communication and Aerospace
<1%

	Technology (ICECA), 2018 Publication	
22	J. Rilling, R. Witte, D. Gasevic, J.Z. Pan. "Semantic Technologies in System Maintenance (STSM 2008)", 2008 16th IEEE International Conference on Program Comprehension, 2008 Publication	<1%
23	Sug Kyun Shin, Ho Hur, Eun Kyung Cheon, Ock Hee Oh, Jeong Seon Lee, Woo Jin Ko, Beom Seok Kim, YoungOk Kwon. "A Personalized and Learning Approach for Identifying Drugs with Adverse Events", Yonsei Medical Journal, 2017 Publication	<1%
24	M A Muslim, A J Herowati, E Sugiharti, B Prasetiyo. "Application of the pessimistic pruning to increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease", Journal of Physics: Conference Series, 2018 Publication	<1%
25	Lecture Notes in Computer Science, 2012. Publication	<1%
26	Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath. "Classification of Sentimental Reviews Using Machine Learning Techniques", Procedia Computer Science, 2015	<1%

	Publication	
27	Desdwyatma Wahyu Wibawa, Muhammad Nasrun, Casi Setianingsih. "Sentiment Analysis on User Satisfaction Level of Cellular Data Service Using the K-Nearest Neighbor (K-NN) Algorithm", 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), 2018 Publication	<1%
28	Mostafa Sayed, Rashed K. Salem, Ayman E. Khder. "A survey of Arabic text classification approaches", International Journal of Computer Applications in Technology, 2019 Publication	<1%
29	Manuel Trenz. "Multichannel Commerce", Springer Nature, 2015 Publication	<1%
30	Vijay Kotu, Bala Deshpande. "Introduction", Elsevier BV, 2015 Publication	<1%
31	"Advances in Data and Information Sciences", Springer Science and Business Media LLC, 2018 Publication	<1%
32	Dimitrios Kravvaris, Katia Lida Kermanidis. "Chapter 20 Opinion Mining for Educational Video Lectures", Springer Nature, 2017 Publication	<1%

33	N. C. Das. "Chapter 2 Decision Complexity and Methods to Meet Them", Springer Science and Business Media LLC, 2015 Publication	<1%
34	Zhu Zhang, Xin Li, Yubo Chen. "Deciphering word-of-mouth in social media", ACM Transactions on Management Information Systems, 2012 Publication	<1%
35	Ayoub Bagheri, Mohamad Saraee, Franciska de Jong. "Sentiment classification in Persian: Introducing a mutual information-based method for feature selection", 2013 21st Iranian Conference on Electrical Engineering (ICEE), 2013 Publication	<1%
36	V. N. Verkhovlyuk, V. A. Morozov, D. V. Stass, A. B. Doktorov, YU. N. Molin. "Experimental and theoretical study of spin evolution ‘switching on’ of the radical ion pair in MARY spectroscopy", Molecular Physics, 2006 Publication	<1%
37	Intelligent Systems Reference Library, 2015. Publication	<1%
38	Y. Huang, B. Q. Huang, M. T. Kechadi. "A new filter feature selection approach for customer churn prediction in telecommunications", 2010 IEEE International Conference on Industrial	<1%

Engineering and Engineering Management, 2010

Publication

39	Anastasia Giachanou, Fabio Crestani. "Like It or Not", ACM Computing Surveys, 2016 Publication	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	Off		