Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 189 (2021) 108–115

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on AI in Computational Linguistics

A Combination of Query Expansion Ranking and GA-SVM for Improving Indonesian Sentiment Classification Performance

Pulung Hendro Prastyo[a]*, Igi Ardiyanto[b], Risanuri Hidayat[c]

[a,b,c]*Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia*

Abstract

The sentiment classification method is a research field that is proliferating in Indonesia since it is fast in extracting public opinion and provides essential and valuable information for stakeholders. Of the best-performing sentiment classification approaches, machine learning is one of them that has excellent performance. However, the method has several problems, such as noisy features and high dimensionality of features that significantly affect the sentiment classification performance. Therefore, to overcome the problems, this paper presents a novel feature selection using a combination of Query Expansion Ranking (QER) and Genetic Algorithm-Support Vector Machine (GA-SVM) for improving sentiment classification performance. Based on the experimental results, the proposed method could significantly improve sentiment classification performance, outperform all state-of-the-art algorithms, and decrease computational time. The method achieved the best performance in average precision, recall, and f-measure with the value of 96.78%, 96.76%, and 96.75%, respectively.

Keywords: Query Expansion Ranking; Genetic Algorithm; Feature Selection; Machine Learning; Sentiment Classification.

1. Introduction

Sentiment classification in Indonesian has emerged as considerable critical attention in the research area. It aims to classify public opinion as positive, neutral, and negative in different fields, such as education, marketing, politics, and economics. It has received much attention because people are more communicative in social media, especially Twitter [1]. A possible explanation for this is the fact that Twitter contains a vast number of tweets or short posts derived from

* Corresponding author. Tel.: +62-0000-000-000;
E-mail address: pulung.hendro@mail.ugm.ac.id

1877-0509
10.1016/j.procs.2021.05.074

its users, allowing people to foster communication by exchanging information and ideas. These tweets are opinions from different people views, including Indonesia which has become one of the countries with the largest active users. Therefore, many Indonesian researchers employed Twitter data as a dataset to extract people's opinions in obtaining important information.

By far, one of the best approaches to classify public opinions is machine learning. It has good accuracy and provides outstanding performance [2,3]. However, the machine learning approach has some challenges in terms of the high dimensionality of feature space and noisy features. Thus, a feature selection is needed to overcome the aforementioned problems since it can select important features, eliminate unnecessary features, minimize the dimensionality of features, and reduce computational time [3–6].

Generally, there are two feature selection approaches, namely filter and wrapper. The use of the filter method is considered more effortless and faster than the wrapper methods. Of the preeminent filter methods, QER [3] is one of them. It provides outstanding performance compared to other filter methods, such as Information Gain (IG), Chi-square, Document Frequency Difference (DFD), and Optimal Orthogonal Centroid (OCFS). However, most filter methods perform no better than the wrapper method [7,8] because they do not evaluate all feature combinations so that there are still noisy features in the classification process. Whereas, wrapper methods evaluate all feature combinations using machine learning. Consequently, those methods have better performance than filter methods although they need high computational cost. Genetic Algorithm (GA) is one of the best wrapper methods that have exceptional performance [6,9–11].

Therefore, this paper attempts to present a novel feature selection using a combination of QER and GA-SVM to overcome the wrapper method's weakness by taking benefit of the filter method for improving sentiment classification. Moreover, omnibus law tweets are chosen as a dataset because this issue has already drawn the attention of Twitter users in Indonesia, and Indonesian tweets are generally written in a mixed language with local or foreign languages that contain slang words [12]. Thus, Indonesian tweets are still open to research.

The remaining part of this paper is organized as follows: Section II commences with related works in the feature selection area. It will then go on to Section III to explain the proposed method offered in this research. After that, the Section IV describes the findings of the research and discussion. Finally, Section V concludes this study and provides insights for future works.

2. Related Works

The feature selection method is an important stage in machine learning-based sentiment classification since it can solve the dimensionality of features and noisy features by selecting important features. Consequently, the method can improve sentiment classification performance. Some feature selection methods have been developed by researchers for sentiment classification, both filter-based and wrapper-based methods.

In the filter-based method, Y. Zhai et al. [13] proposed Chi-square as a feature selection. They compared Chi-square with Information Gain (IG). Based on their experimental results, they found that Chi-square could improve the classification performance and outperformed IG. Moreover, other studies [14,15] reported that chi-square could increase the classification method's accuracy and was more effective in the computational cost.

A. S. Manek et al. [16] presented a Gini Index (GI) as a feature selection method with SVM for sentiment classification. Their proposed method was compared with other feature selection methods, such as Maximum Relevance, Correlation, and IG on movie reviews. According to the results, their proposed method had better classification performance than other methods in terms of reduced error rate and accuracy using a large movie review dataset.

Mihuandayani et al. [17] employed IG feature selection to select the relevant feature to the tax topic and used SVM as a classification algorithm to classify tax service relied on public opinion. The dataset was obtained from Facebook and Twitter. Based on positive and negative sentiment, they classified data into three categories: service, website, and news. They stated that their research could be used as a basis for big data analysis in tax cases based on public opinion. Moreover, this study [18] indicated that IG could increase the run-time efficiency of the Naïve Bayes classifier. However, their research still has shortcomings, namely several misclassifications in which their method cannot classify certain sentiments well.

Other scholars, such as B. V. G. Bispo and T. N. Rios [19], presented a novel feature selection algorithm for text classification called Statera. Nine real document collections with different characteristics were used as the dataset in their study. Their experiments concluded that the Statera using the Naïve Bayes (NB) classifier was superior to Mutual Information (MI) and the Chi-square method. Additionally, their proposed method can outperform state-of-art feature selection techniques.

T. Parlar et al. [3] presented a new feature selection called Query Expansion Ranking (QER) to assess the necessary words for query expansion. They compared their approach to other feature selection methods such as IG, Chi-square, DFD, and OCFS. Turkish and English review datasets were used in their experiments. The experimental results indicated that QER increased sentiment analysis performance in terms of classification accuracy and computational cost.

In the wrapper methods, R. Shahid et al. [2] presented a Biogeography Based Optimization (BBO) algorithm to select optimal features. They classified product reviews from www.amazon.com using SVM and NB algorithms. They also used accuracy as the evaluation measure. In the experimental results, they stated that BBO could improve the accuracy of SVM and reduce the number of features. However, the accuracy that they have improved was still low.

Another existing research by S. Ernawati et al. [10] implemented the NB algorithm with the GA as a wrapper feature selection to analyze sentiment reviews of online fashion companies. They employed NB as a classifier to classify text sentiment into positive and negative sentiment; meanwhile, GA was used to select the best features in improving the performance of the NB algorithm. Moreover, customer reviews of online fashion companies were selected as a dataset. Accuracy and AUC were employed as the evaluation measures. It is interesting to note that GA could significantly increase the accuracy and AUC of classifiers according to the experimental results. In other papers [6,9–11], GA has exceedingly improved machine learning algorithms' efficiency.

Taken together, these previous studies that have been mentioned before support the ideas that filter methods are more effective in the computational cost. However, filter methods still present a hindrance in terms of noisy features. Meanwhile, most wrapper methods can improve sentiment classification performance, but those methods undoubtedly require high computational time because they investigate all features to achieve the best feature combination. This study [8] supported the statements as the researchers stated that the wrapper method is better than the filter method in terms of accuracy.

Therefore, the researchers proposed a novel feature selection using a combination of QER and GA-SVM. It is hoped that the work presented here provides some insights to overcome the wrapper method's disadvantage by taking the benefit of the filter method to improve sentiment classification performance and decrease the computational time.

3. Proposed Methods

This stage explains data collection, preprocessing, feature selection, and the machine learning algorithm used in the experimental setup. The framework of the proposed method is shown in Fig. 1.

3.1. Dataset

The researchers obtained 14,097 tweets on Twitter using the GetOldTweets3 library from 8 July to 29 July 2020. GetOldTweets3 uses sentence, words, or hashtag keywords to retrieve tweets. In this study, #omnibuslaw and omnibus law were used as keywords. After that, duplicated and unused data were eliminated. The researchers employed 4,000 labeled tweets for the training process, where 2,000 of them were categorized as a positive sentiment and the other 2,000 tweets were considered as a negative sentiment.

3.2. Preprocessing data

The tweets gathered from Twitter do not follow standard words and not structured well. Subsequently, several pre-processing steps are then used to improve tweet data quality, namely:

- Removing Twitter special symbols, punctuation, number, URLs, and other characters.

- Case folding, which refers to the process of converting all the characters in a document into lowercase letters. Lower function in python was used to change tweets into lowercase letters.
- Slang word, that is generally processed by changing non-standard words into standard words. In this study, 3,048 words were used to handle Indonesian slang word problems.
- Stemming, whose goal is to minimize some types of words to the root form. Sastrawi library [20] was used in this study.
- And Tokenization, which is a process of separating tweets into words. It was done by using Natural Language Toolkit (NLTK) library in python.

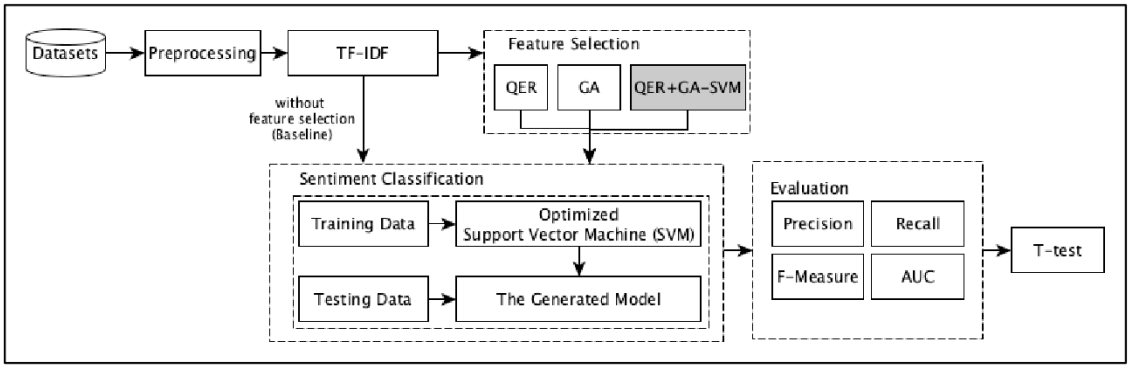

Fig. 1. The proposed framework.

3.3. Feature extraction

Term Frequency-Inverse Different Frequency (TF-IDF) was used as feature extraction in this research because TF-IDF is more accurate than Bag of Word (BoW) [21]. TF-IDF converts texts into vectors so that machine learning can process the tweets well. TF calculates the frequency at which a word appears in a tweet. Meanwhile, IDF measures how significant a word is. It does not accept words occurring several times. In particular, the equation of $tf - idf = tf_t \times idf_t$, where tf_t is the term frequency and idf_t is the inverse document frequency.

3.4. Integration of QER and GA-SVM

In this study, QER was selected as a filter-based feature selection because it performed better than IG, Chi-square, DFD, and OCFS [3]. QER was used to reduce and select the best features in the initial selection. The features of the dataset were then divided into 8 thresholds from 5,468 features (without feature selection), namely 1,500, 2,000, 2,500, 3,000, 3,500, 4,000, 4,500, and 5,000 features. After that, the selected features in the form of a sparse matrix were utilized as an input to GA-SVM. Finally, GA-SVM optimized the selected features to obtain the most optimal features because the QER selection process still had irrelevant features. QER was computed using Eq. (1).

$$Score_f = \frac{p_f + q_f}{p_f - q_f} \tag{1}$$

In this equation, $Score_f$ denotes the QER value. Then, p_f is the ratio of positive documents that contain feature f and q_f is the ratio of negative documents that contain feature f. The value of p_f and q_f can be calculated using Eq. (2) and Eq. (3).

$$p_f = \frac{df_+^f + 0.5}{n^+ + 1.0} \tag{2}$$

$$q_f = \frac{df_-^f + 0.5}{n^- + 0.5} \quad (3)$$

Here, df_+^f is the total number of positive documents that contain feature f. df_-^f is the total number of negative documents that contain feature f and n^+ is the number of positive documents. Finally, n^- is the number of negative documents.

GA is a wrapper-based feature selection that interacts directly with the machine learning algorithm to evaluate all features. The purpose of GA is to find optimal solution set. Each solution set is called a population which consists of vectors such as chromosome. Each item in chromosome is called a gene. In this proposed method, chromosomes represent features, which are encoded as binary string of 1 and 0. In this case, 1 means that the features are selected and 0 indicates that the features are removed [22].

Moreover, this research combined GA and SVM to select the best features. SVM was employed as a classifier to get an f-measure that is an input value in GA's fitness. One of the challenges in the SVM is determining the optimal parameters so that the SVM's parameters were optimized using GridSearchCV to get optimal parameters provided by the scikit-learn library (called optimized-SVM). Further, the fitness value was determined using Eq. (4) to minimize the number of used features and maximize model performance [23]. After that, data were divided into two, consisting 70% of the training data and the other 30% of testing data. The data were then validated using 5-fold cross-validation.

Finally, GA procedures, such as generating initial chromosomes, parent selection, crossover process, and mutation process, were run to discover the best solution until the stopping criteria (generation) were satisfied. The smaller the chromosome's fitness value, the better the proposed method performance. In this scheme, the smallest chromosome's fitness value would be used as features in the sentiment classification stage. The steps of the proposed method can be seen in Fig. 2.

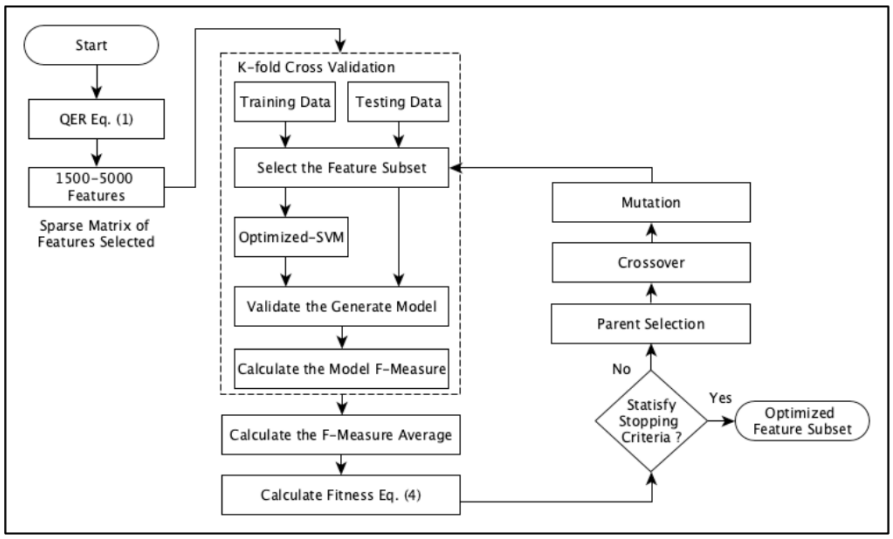

Fig. 2. The proposed feature selection.

$$Fitness_{GA} = a(1-P) + (1-a)\left(1 - \frac{N_f}{N_t}\right) \quad (4)$$

Where P denotes the f-measure value of the optimized-SVM algorithm; N_f is the size of feature subset tested; and N_t is the total number of features. The term on the left side of the equation describes the overall accuracy of the model, while the term on the right explains the percentage of features used. Constant $a \in [0,1]$ states the weight of the study objective, performance, and subset size. In this research work, a was given a weight of 0.88.

The optimal parameter selection is a fundamental property in GA as it can directly affect GA. There are four meaningful parameters in GA, namely generation (gen), population (pop), crossover probability (pc), and mutation probability (pm). Therefore, this study conducted twenty-seven (27) experiments on these parameters to obtain optimal

parameters. Generation was tested using the values of 20, 60, and 100, respectively. After that, the population was assigned the value of 10, 30, and 50. Then, crossover probability used the value of 0.5, 0.6, and 0.7. Finally, mutation probability was given the value of 0.001.

3.5. Machine learning algorithm

In this study, Support Vector Machine (SVM) is a machine learning algorithm employed to classify the sentiment since it provided competitive performances [24–27]. However, there are two challenges in SVM, such as optimal parameter selection and kernel function selection. Thus, GridSearchCV was used to optimize SVM's parameters according to the problems. At the same time, Radial Basis Function (RBF) kernel SVM was employed as kernel function because it is suitable for sentiment classification problems. Moreover, its performance is better than other kernels [28,29]. Choosing the right kernel function gives maximum results and vice versa.

3.6. Evaluation and validation

In the sentiment classification process, 5-fold cross-validation was used to validate the SVM on the omnibus law dataset. The training and testing data were then divided into 70% and 30%, respectively. Precision, recall, and f-measure were employed to evaluate the proposed method because they have become the standard method. Besides, AUC was also employed to understand the algorithm's performance criteria. The AUC criteria are described as follows [30]:

- 0.90 - 1.00 = excellent classification,
- 0.80 - 0.90 = good classification,
- 0.70 - 0.80 = fair classification,
- 0.60 - 0.70 = poor classification,
- 0.50 - 0.60 = failure.

4. Results and Discussion

Firstly, this study compared the sentiment classification results of the proposed method (QER+GA-SVM) with the baseline algorithm (without feature selection), QER, and GA. As shown in Fig. 3, the proposed method outperformed all state-of-the-art algorithms. The differences in the value of precision, recall, f-measure, and AUC were 1.68%, 1.67%, 1.67%, and 0.23%, respectively, against the baseline algorithm. The proposed method achieved better performance than the previous studies, including GA [6,9–11] and QER [3], which obtained excellent performance compared to IG, Chi-square, DFD, and OCFS. Moreover, the AUC score of the proposed method was classified as an excellent classification.

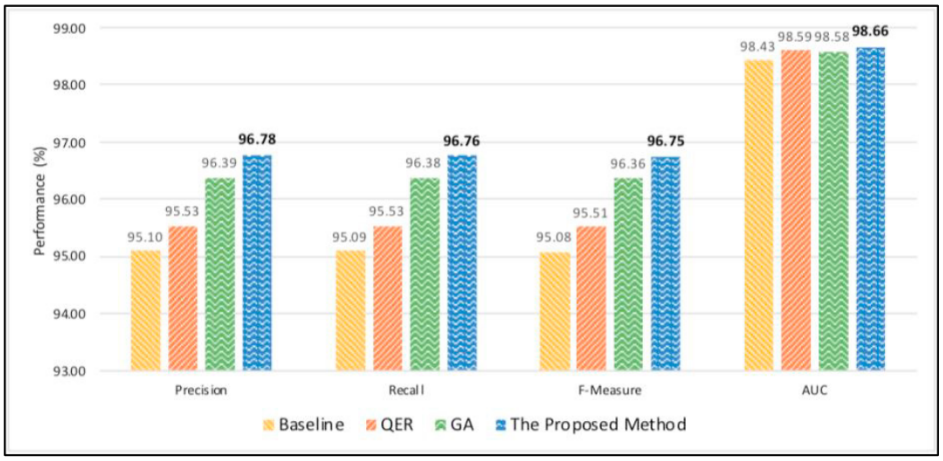

Fig. 3. The comparison of computational time of each algorithm.

Secondly, this study also compared the computational time of the baseline algorithm with the feature selection methods. For data testing (using selected features), the proposed method requires 25.4 seconds faster than the baseline algorithm, 7.89 seconds faster than QER, and 2.47 seconds faster than GA. It indicates that the proposed method could improve sentiment classification performance and decrease computational time. The results can be seen in Table. 1

Table 1. The comparison of computational time of each algorithm

Algorithm	Computational Time (Seconds)
Baseline	31.75
Query Expansion Ranking [3]	14.24
Genetic Algorithm [6,9–11]	8.82
The Proposed Method	**6.35**

Lastly, this research work employed a statistical t-test (paired two sample for means) to verify a significant difference between the proposed method and the baseline algorithm [22]. This study set the statistical significance level (α) to 0.05. It implies that there is a statistically significant difference if the p-value is less than 0.05. Based on the t-test result, the proposed method could significantly improve the sentiment classification performance, which the proposed method obtained a p-value of 0.0001 (p-value < 0.05).

5. Conclusion

This study has presented a novel feature selection using a combination of Query Expansion Ranking (QER) and Genetic Algorithm-Support Vector Machine (GA-SVM). Based on the experimental results, our proposed method could significantly improve the sentiment classification performance. The proposed method outperformed all state-of-the-art methods by achieving the precision, recall, and f-measure value of 96.78%, 96.76%, and 96.75%, respectively. The obtained AUC score in this study was 98.66%, which meant the proposed method was classified as an excellent classification. Moreover, the proposed method could also decrease the computational time from the baseline algorithm. In short, robust performances are the most significant findings to emerge from this study.

The issue of other filter and wrapper methods is an intriguing area for further research to achieve different results and more definitive evidence. Moreover, continued efforts of the researchers are necessitated in utilizing more experiments and determining the optimal parameters of GA.

Acknowledgements

This research work is supported by The Indonesia Endowment Fund for Education, Lembaga Pengelola Dana Pendidikan (LPDP).

References

[1] Shelke PP, Korde AN. Support Vector Machine based Word Embedding and Feature Reduction for Sentiment Analysis-A Study. Proc. Fourth Int. Conf. Comput. Methodol. Commun. (ICCMC 2020), 2020, p. 176–9.

[2] Shahid R, Javed ST, Zafar K. Feature Selection Based Classification of Sentiment Analysis using Biogeography Optimization Algorithm. 2017 Int. Conf. Innov. Electr. Eng. Comput. Technol., IEEE; 2017, p. 1–5. https://doi.org/10.1109/ICIEECT.2017.7916549.

[3] Parlar T, Özel SA, Song F. QER : a new feature selection method for sentiment analysis. Human-Centric Comput Inf Sci 2018;8:1–19. https://doi.org/10.1186/s13673-018-0135-8.

[4] Zeng D, Peng J, Fong S, Qiu Y, Wong R. Medical data mining in sentiment analysis based on optimized swarm search feature selection. Australas Phys Eng Sci Med 2018;41:1087–100. https://doi.org/10.1007/s13246-018-0674-3.

[5] Tian W, Li J, Li H. A Method of Feature Selection Based on Word2Vec in Text Categorization. 2018 37th Chinese Control Conf., Technical Committee on Control Theory, Chinese Association of Automation; 2018, p. 9452–5.

[6] Bidi N, Elberrichi Z. Feature Selection For Text Classification Using Genetic Algorithms. 2016 8th Int. Conf. Model. Identif. Control, University of MEDEA, Algeria; 2016, p. 806–10. https://doi.org/10.1109/ICMIC.2016.7804223.

[7] Kurniawati I, Pardede HF. Hybrid Method of Information Gain and Particle Swarm Optimization for Selection of Features of SVM-Based Sentiment Analysis. 2018 Int Conf Inf Technol Syst Innov ICITSI 2018 - Proc 2019:1–5. https://doi.org/10.1109/ICITSI.2018.8695953.

[8] Gokalp O, Tasci E, Ugur A. A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification. Expert Syst Appl 2020;146:1–10. https://doi.org/10.1016/j.eswa.2020.113176.

[9] Muthia DA, Putri DA, Rachmi H, Surniandari A. Implementation of Text Mining in Predicting Consumer Interest on Digital Camera Products. 2018 6th Int. Conf. Cyber IT Serv. Manag., IEEE; 2018, p. 1–7. https://doi.org/10.1109/CITSM.2018.8674063.

[10] Ernawati S, Yulia ER, Frieyadie, Samudi. Implementation of the Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies. 2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018, IEEE; 2018, p. 6–10. https://doi.org/10.1109/CITSM.2018.8674286.

[11] Aliane AA, Aliane H, Ziane M, Bensaou N. A Genetic Algorithm Feature Selection Based Approach for Arabic Sentiment Classification. 2016 IEEE/ACS 13th Int. Conf. Comput. Syst. Appl., IEEE; 2016, p. 1–6. https://doi.org/10.1109/AICCSA.2016.7945661.

[12] Hidayatullah AF. Language tweet characteristics of Indonesian citizens. Proc. 2015 Int. Conf. Sci. Technol. TICST 2015, IEEE; 2015, p. 397–401. https://doi.org/10.1109/TICST.2015.7369393.

[13] Zhai Y, Song W, Liu X, Liu L, Zhao X. A Chi-square Statistics Based Feature Selection Method in Text Classification. 2018 IEEE 9th Int. Conf. Softw. Eng. Serv. Sci., IEEE; 2018, p. 160–3.

[14] Haryanto AW, Mawardi EK. Influence of Word Normalization and Chi-squared Feature Selection on Support Vector Machine (SVM) Text Classification. 2018 Int Semin Appl Technol Inf Commun 2018:229–33.

[15] Nurhayati, Putra AE, Wardhani LK, Busiman. Chi-Square Feature Selection Effect On Naive Bayes Classifier Algorithm Performance For Sentiment Analysis Document. 2019 7th Int. Conf. Cyber IT Serv. Manag., 2019, p. 1–7.

[16] Manek AS, Shenoy PD, Mohan MC, R VK. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World Wide Web 2017;20:135–54. https://doi.org/10.1007/s11280-015-0381-x.

[17] Mihuandayani, Utami E, Luthfi ET. Text mining based on tax comments as big data analysis using SVM and feature selection. 2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018, IEEE; 2018, p. 537–42. https://doi.org/10.1109/ICOIACT.2018.8350743.

[18] Widya Sihwi S, Prasetya Jati I, Anggrainingsih R. Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier. Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018, IEEE; 2018, p. 190–5. https://doi.org/10.1109/ISEMANTIC.2018.8549757.

[19] Bispo BVG, Rios TN. Statera : A Balanced Feature Selection Method for Text Classification. 2018 7th Brazilian Conf. Intell. Syst., 2018, p. 260–5. https://doi.org/10.1109/BRACIS.2018.00052.

[20] Sastrawi. Sastrawi Library 2020. https://pypi.org/project/Sastrawi/ (accessed November 5, 2020).

[21] Alzami F, Udayanti ED, Prabowo DP, Megantara RA. Document Preprocessing with TF-IDF to Improve the Polarity Classification Performance of Unstructured Sentiment Analysis. Kinet Game Technol Inf Syst Comput Network, Comput Electron Control 2020;5:235–41. https://doi.org/10.22219/kinetik.v5i3.1066.

[22] Wahono RS, Suryana N, Ahmad S. Metaheuristic Optimization based Feature Selection for Software Defect Prediction. J Softw 2014;9. https://doi.org/10.4304/jsw.9.5.1324-1333.

[23] Vieira SM, Mendonc LF. Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. Appl Soft Comput 2013;13:3494–504. https://doi.org/10.1016/j.asoc.2013.03.021.

[24] Prastyo PH, Sumi AS, Dian AW, Permanasari AE. Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel. J Inf Syst Eng Bus Intell 2020;6:112–22. https://doi.org/10.20473/jisebi.6.2.112-122.

[25] Banik N, Hasan Hafizur Rahman M. Evaluation of Naïve Bayes and Support Vector Machines on Bangla Textual Movie Reviews. 2018 Int. Conf. Bangla Speech Lang. Process. ICBSLP 2018, IEEE; 2018, p. 1–6. https://doi.org/10.1109/ICBSLP.2018.8554497.

[26] Rahat AM, Kahir A, Masum AKM. Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset. 8th Int. Conf. Syst. Model. Adv. Res. Trends, 2019, p. 266–70. https://doi.org/10.1109/smart46866.2019.9117512.

[27] Shamantha Rai B, Shetty SM, Rai P. Sentiment analysis using Machine learning classifiers: Evaluation of performance. 2019 IEEE 4th Int. Conf. Comput. Commun. Syst. ICCCS 2019, IEEE; 2019, p. 21–5. https://doi.org/10.1109/CCOMS.2019.8821650.

[28] Prastyo PH, Ardiyanto I, Hidayat R. Indonesian Sentiment Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF. 2020 Int. Conf. Data Anal. Bus. Ind., 2020, p. 1–6. https://doi.org/10.1109/icdabi51230.2020.9325685.

[29] S. S, K.V. P. Sentiment analysis of malayalam tweets using machine learning techniques. ICT Express 2020;6:300–5. https://doi.org/10.1016/j.icte.2020.04.003.

[30] Gorunescu F. Data Mining: Concepts, Models and Techniques. Volume 12. Springer-Verlag Berlin Heidelberg; 2011.