Opinion mining on culinary food customer satisfaction using naïve bayes based-on hybrid feature selection

Oman Somantri, Dyah Apriliani
Department of Informatics, Politeknik Harapan Bersama Tegal, Indonesia

Article Info

Article history:

Received Nov 26, 2018
Revised Jan 21, 2019
Accepted Mar 1, 2019

Keywords:

Culinary food
Customer satisfaction
Hybrid feature selection
Naïve bayes
Opinion mining

ABSTRACT

Conducting an assessment of consumer sentiments taken from social media in assessing a culinary food gives useful information for everyone who wants to get this information especially for migrants and tourists, in th other hand that information is very valuable for food stall and restaurant owners as information in improvinf food quality. Overcoming this problem, a sentiment analysis classification model using naïve bayes algorithm (NB) was applied to get this information. This problem occurs is the level of accuracy of classification of consumer ratings of culinary food is still not optimal because the weight of values in the data preprocessing process are not optimal. In this paper proposed a hybrid feature selection models to overcome the problems in the process of selecting the feature attributes that have not been optimal by using a combination of information gain (IG) and genetic algorithm (GA) algorithms. The result of this research showed that after the experiment and compared to using others algorithms produce the best of the level occuracy is 93%.

Corresponding Author:

Oman Somantri,
Department of Informatics,
Politeknik Harapan Bersama Tegal,
Jln.Mataram No.09 Pesurungan Lor, Kota Tegal 52141, Indonesia.
Email: oman.somantri@poltektegal.ac.id

1. INTRODUCTION

Social media has a major influence to the development of information media where to get the origin of the information is difficult, but using the media finally the desired information can be easily obtained and more quicly, in a matter of hours and minutes [1]. There are various social media that are often used by many people, such as a blogs, twitter, facebook, youtube, tripadvisor, instagram and others [2]. Through this media opinions emerged from each individual which contained subjective assessments about various things, one of them was an assessment of food and culinary attraction. Culinary food is particular attraction for everyone, using social media nowadays many people make social media a benchmark in assessing a culinary food so that it becomes a decision supporter to try these food. Through rhe comments and people opinions who have experienced culinary food at place that has been visited, it can be used as a support for decisions of customers who in this culinary connoisseurs to come to the place as well as supporting the decision of the stall owners and culinary restaurants to be made as reference for the success rate of the form of service to its customersv [3], beside that it can also be used as aa media for tourism promotion for tourists and imigrants [4].

The problem that occurs, sometimes customers can not read comments too much to get a recomended decision the right choice, in the other hand the problems occur of food stall and restaurant owner who want to obtain data about comments from culinary connoisseurs to their place to be able to determine a decision related to service given according to the wishes of the customers or still need an increase in service, maybe in terms of food, comfortable place or service at that place. Related to the existing problems, a method is needed that can help to analyze the related comments. The solution is the implementation of a sentiment analysis model (SA) or opinion mining in which using dataset from social media becomes a decision supporter [5].

Nowadays Analytical sentiment is applies to many research object, as film reviews [6, 7], food reviews, certain product reviews [8], tourist attractions [9], hotel reviews [10, 11] etc. Analytical Sentiment (AS) is a part of computer science, which is works through a process of understanding and then extracting and processing textual dataset automatically [12, 13]. AS works to get information sentiment contained in it an opinion sentence that is subjective assesment [14]. Nowadays AS working to see tendencies from opinions that is a problem or object carried out by someone lead to a positive or negative opinios and it can be that the opinion that emerges is neutral so that it becomes a decision support material. As a part of the science text mining, AS is widely used to classify an example of the data text from various sources as short text, example short stories, abstract text, news, articles, website informations, etc [15]. Nowadays there are technics learning machine methode has been used, as Neural Network, Support Vector Machine, Naive Bayes, Decision Tree, k-Nearest Neighbours and Bayesian Network [16].

In this paper, the sentiment of satisfication assesment for culinary food is applied to a classification algorithm that is Naive Bayes (NB). Naive Bayes (NB) is a one of good algorithm in classification analysis sentiment analysis [17-19]. On the other hand, there are problems that occur, namely the existence of problems in the data processing process, one of which is the weight of process. The selection of appropriate weight value is one of the keys in the learning process carried out by the algorithm used so that it influences the level of accuracy produced. Some research related optimalizing the level of accuracy classification in text mining previously carried out by researchers, among others, by the process feature selection using several optimization algorithm. Some researchers caried out optimization to overcome these problems using Particle Swarm Optimization (PSO) [20–24], genetic algorithm (GA) [25-27], information gain (IG) [28], Gini index [29] and other algorithms. In the right side this solution is not enough tom provide a significant level of accuracy and the need for optimation.

Based on the research that has been done before, this paper purposes a model with a feature selection by applying 2 optimization algorithms, are Information Gain and Genetics algorithm which implementation on Naive Bayes model so that the level accuracy of classification on sentimen assesment satisfication and presentation more high presentation results.

2. PROPOSED METHOD

Proposed method in this paper is a hybrid feature selection model that using analysis sentiment for customers satisfication assesment of culinary food. Proposed model are integrated between Information Gain and Genetic Algorithm, to get the best result so the implementation on SVM algorithm, NB, k-NN and Decision Tree. Evaluation of proposed model as seen at the Figure 1, carried out with comparing of models that after has been applied IG and GA models. Sistem validation uses Cross k-Fold validation, which is expected to show the best level of accuracy on the existing model [30]. Dataset in this research divided in two part, they are data training and data testing, where data training used to get the expected model, while data testing used to testing the dataset, is the model obtain as expected. The final result of this study is to obtain the best model with the highest performance and has the highest presentation.

2.1. Information Gain (IG)

In this paper, IG apply as metode on the model. This is need to carried out because IS is one of the best algorithm that can be use for feature selection [28]. Calculation of *Information Gaint* is done by using equation:

$$Info\ (D) = -\sum_{i=1}^{c} \quad p_i log_2(p_i) \tag{1}$$

With:
c: number of values in the target atribut (number of classification classes)
pi: number of sample for class i

$$Info_A(D) = \sum_{j=1}^{V} \left(\frac{Dj}{D}\right) x\ Info\ (D_j) \tag{2}$$

For measure the effectiveness of an attribute in classifying the data calculated by equation :

$$\text{Gain(A) = | Info(D) – Info}_\text{A}\text{(D)} \tag{3}$$

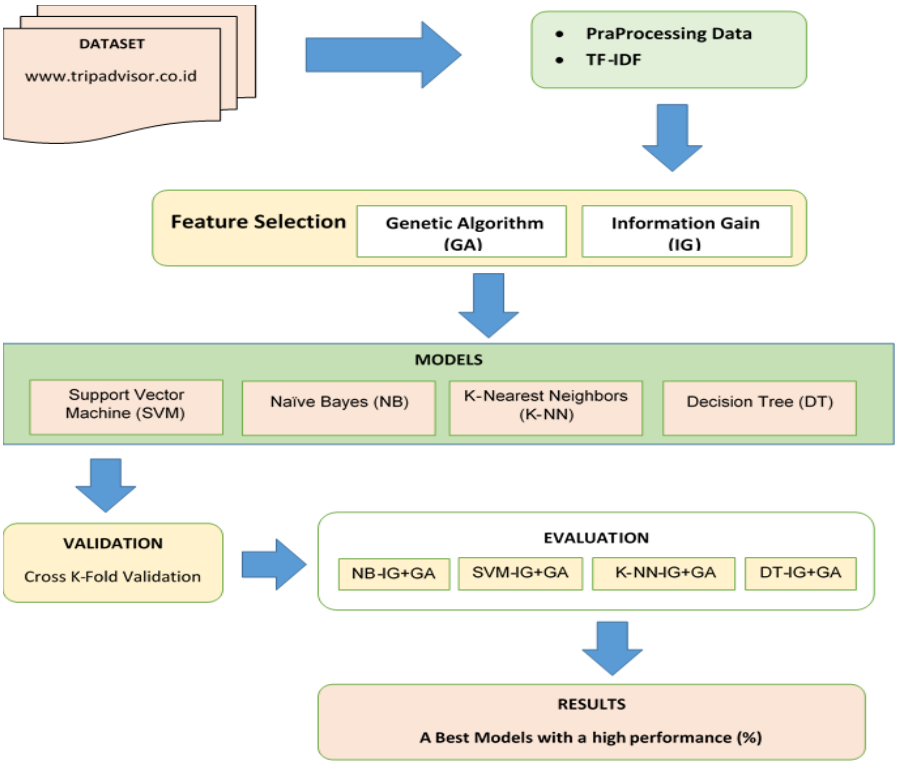

Figure 1. Proposed method (Hybrid IG and GA)

3.2. Genetic Algorithm

Genetic Algorithm (GA) is a searching methode whose work patterns are based on the principle of genetic process and natural selection. Search processing that done by GA is appropriate with genetic process from biology organisms that based on a evolutionary theory [31]. This algorithm used to be alternative on decision of a feature selection in order to get the model that is optimal model [32].

3. RESEARCH METHOD

The Proposed model in this study is an area consisting of several methods which integrated into an algorithm that has been determined based on its capabilities and strength. The steps taken to get the best model consist of several stages, including the process of processing the document process data which consist of the weighting tokenisation filtering understanding of the attributes and aplication of the model adn the las part is data validation.

3.1. Dataset and Materials

The first stage carried out in the sentiment model analysis of the assesment of customer satisfication in culinary food is the process of collecting dataset. Dataset used in this experiment are data taken one of sites www.tripadvisor.com taken during the period of data collection in 2017 and 2018. Dataset taken is the text of opinions written by the site visitors on food and culinary stalls found in Tegal city, Indonesia. In the process collecting this dataset, dataset used is limited only to Indonesian text data.

3.2. PraProcessing Data

In this study, the preprocessing data carried out to get input data that appropriate with proposed model. In this process doing by some steps, one of them is tokenized where in this process done by separatation of text data, that separated by each syllable with a space separator. The next step is done by transformcase, in this step, the existing text data is changed all into lowercase text data with minimal char is=4 and maximal char is=20. On this process, the class is alrady done, that is, displaying word that have entered you already well or not into the data training model tjat has been prepared, of course, the data used for stopword is Indonesian words. In this step doesnot do process Stemming, that is change every word which consist prefix and suffix are basic words.

3.3. Weighting TF-IDF

This step carried out to get a weight value obtained on each feature. At this steps we give weighting using term patterns frequency or amount term in every documents, and inverse document frequency or invers from amount documents in the term. Weigthing process in every term in this step use Term Frequency-Invers Document Frequency (TF-IDF) method [33].

$$W_{i,j} = tf_{i,j}\ x \log(\frac{N}{dfi}) \tag{4}$$

where,
tf_{ij} = number of occurences of $_i$ in $_j$
df_i = number of documents containing $_i$
N = total number of documents

4. RESULTS AND DISCUSSION

To get the best experimental result need hardware and software that accordance with is expected, this research use software Rapidminer with operating system windows 7, processor system Intel Core i%, and 4 GB memory. Experiment were carried out by applying proposed model into seveal algorithm including Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbors (K-NN), dan Decision Tree (DT). Experiment was carried out using a hybrid model proposed, namely feature selection using Information Gain (IG) and Genetic Algorithm (GA).

4.1. Classical Model

The result in this hybryd model, Information Gain (IG) is combaining with several algorithm model namely Support Vector Machine (SVM-IG), Naïve Bayes (NB-IG), K-Nearest Neighbors (k-NN-IG), dan Decision Tree (DT-IG). In this experiment several models were produced which had defferent levels of accuracy from each other. At Table 1, can be seen result the model which get by using several algorithm model frequently used. On this model using two validation model, they are k-Fold 10 and k-Fold 5, so that it can show the differences with the accuracy some existing models.

Table 1. The Result Comparing Accuracy Algorithm

Model	Validation	
	k-Fold=10	k-Fold=5
SVM	69.36%	69.17%
Decision Tree	74.87%	73.29%
Naïve Bayes	68.72%	73.26%

Table 2. The Result Accuracy SVM-IG+GA

kernel	Validation				
	k-Fold=10	k-Fold=8	k-Fold=6	k-Fold=4	k-Fold=2
dot	74.23%	74.90%	75.61%	74.82%	77.21%
radial	77.12%	77.19%	78.03%	77.92%	74.81%
polynomial	74.68%	73.33%	74.78%	74.09%	74.80%

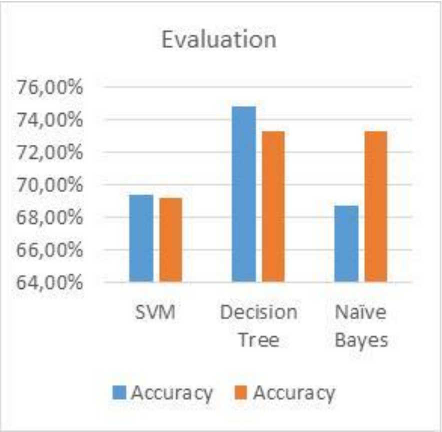

Figure 2. Result of comparative model

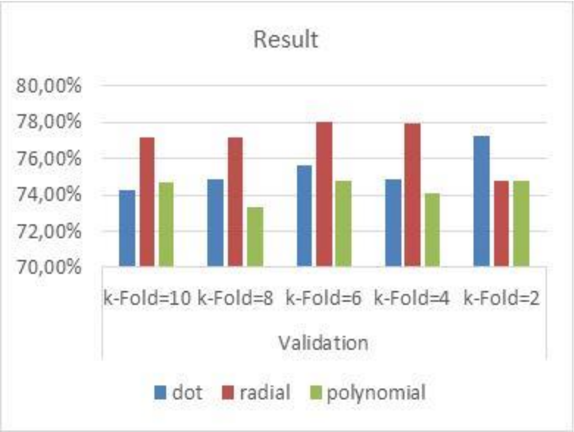

Figure 3. Effects of selecting different switching under dynamic condition

At Table 1 shows the level accuracy from every existing model and have different result. SVM model has the highest accuracy 69,36% by using k-Fold 5, but Decision Tree has the level accuracy 74.87%. Different from that produced by Naive Bayes, this model has level accuracy 73,26% almost same as its accuracy with Decision Tree model. The visual description of the result of analysis model shown at Figure 2, shows that the highest is generated by the Decision Tree model.

4.2. Support Vector Machine (SVM) and Hybrid Model

Proposed model in this paper are hybryd feature selection model that applied on existing dataset, namely IG and GA. To obtain another level accuracy value, experiment of feature selection was carried out using GA. Process GA was applied on several existing model, as SVM, Naive Bayes, K-NN and Decision Tree. On this stage IG still used, expected can be increase the level accuracy. For first stage carried out to applied experiment GA into SVM and IG or can be namely model (SVM-IG+GA), and the result can be shown in Table 2. The experimental result show, that SVM by applied hybrid IG and GA model in Table 2, the highest accuracy level is 78,03%. Result of The best model was applied using k-Fold 6 and kernel radial type. In the other side, shown the result of highest accuracy level was applied dot kernel type with the accuracy was 77.21% adn k-Fold 2. Further, shown at Table 2 for SVM by using polynomial kernel type has result the highest level accuracy was 74.80% with k-Fold 2. Based on the result from the experiment was get description shown at Figure 3.

4.3. Naïve Bayes (NB) and Hybrid Model

The result experimental were using Naive Bayes algorithm (NB) by applied combination IG and GA (NB-IG+GA). At this model, using GA was a part of feature selection in order to get value with the best level accuracy. The combination result was shown at Table 3, shown was the proposed model incresing a good level accuracy.

Table 3. The accuracy result NB-IG+GA

Sampling	Validation				
	k-Fold=10	k-Fold=8	k-Fold=6	k-Fold=4	k-Fold=2
linear	84.94%	85.16%	**92.93%**	89.69%	67.47%
shuffled	78.14%	77.14%	77.89%	76.46%	77.19%
statified	76.47%	77.86%	75.54%	77.92%	77.99%

At Table 3 show the NB-IG+GA model, the level of accuracy produces has increased significantly. The accuracy result using k-Fold 6 had the best level accuracy, was 92.93% by using linear sampling. At Table 3 shown the result accuracy by using shuffled sampling and stratified still lower than linear sampling, if shown the result still low. In detail the model results are displayed at Figure 4. Figure 4 shown the model NB-IG+GA by using linear sampling has the level accuracy more better than others, this is seen even though the lowest value of accuracy obtained on k-Fold 2 is 67.47% but on the other hand obtains the highest level of accuracy compared to other models.

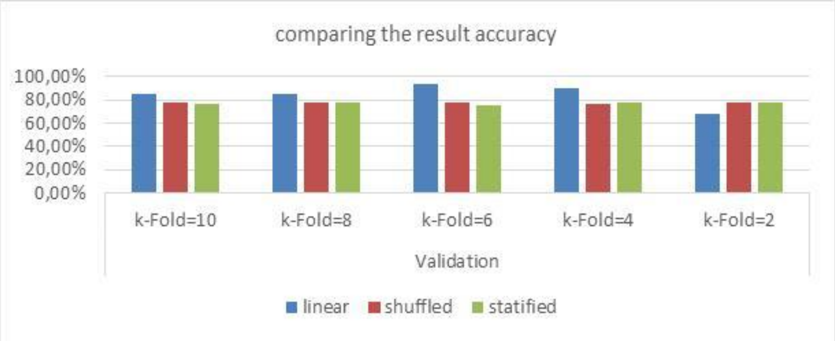

Figure 4. Comparing the result accuracy by using Naive Bayes

4.4. K-Nearest Neighbors (K-NN) and Hybrid Model

The next experiment is to apply the GA feature selection by using IG on K-NN algorithm. In this model, hybryd IG+GA model into K-NN showing the experimental result in Table 4.

The result accuracy obtains as shown in Table 4 and Figure 5, shows the highest level of accuracy using parameters k(optimal)=1, where using linear sampling produced an accuracy 74,10% adn 77.88% by suffed sampling. Different from the result has been applied by statified sampling, the highest level accuracy has been obtain 77.18% but by using k (optimal)=2.

Table 4. The accuracy result KNN-IG+GA

k	Sample		
(Optimal)	linear	shuffled	statified
k=1	74.10%	77.88%	75.00%
k=2	73.97%	76.35%	77.18%
k=3	73.91%	77.12%	76.54%
k=4	72.31%	76.28%	75.83%
k=5	73.14%	77.37%	77.37%
k=6	72.44%	76.54%	76.35%
k=7	72.50%	75.90%	75.64%
k=8	70.90%	75.58%	75.77%
k=9	70.83%	74.94%	74.17%
K=10	68.46%	74.94%	74.94%

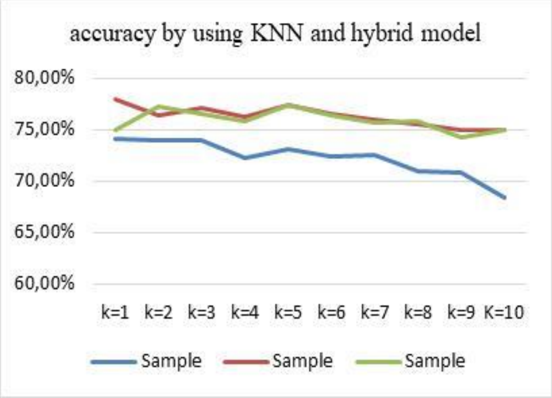

Figure 5. The result accuracy by using K-NN & hybrid Model

Table 5. The accuracy result by Decision Tree with IG+GA

Criterion	Sample		
	linear	shuffled	statified
gain_ratio	76.15%	77.31%	77.37%
information_gain	74.68%	72.31%	73.27%
gini_index	73.91%	74.94%	74.87%
accuracy	73.14%	77.12%	77.31%

Table 6. Result of IG Model dengan GA

Model	Accuracy
SVM-IG+GA	78.03%
NB-IG+GA	92.93%
K-NN-IG+GA	77.88%
Decision Tree-IG+GA	77.37%

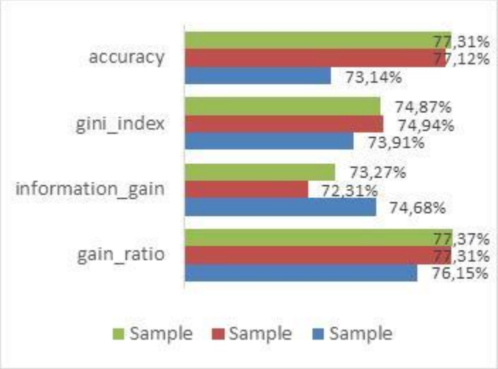

Figure 6. The accuracy result by DT-IG+GA

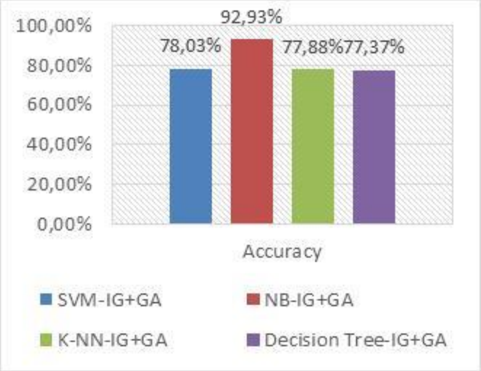

Figure 7. The results level of accuracy in each model

4.5. Decision Tree (DT) and Hybrid Model

Based on Decison Tree model algorithm feature selection apply IG and GA, the experiment carried out at seen in Table 5. Criterion paramaters has been used on DT produced differen level of accuracy, so that the right selection paramater must be done. In Table 5 can be seen the accuracy result by combaining based on criterion paramaters with existing model. By using gain_ratio paramaters, the highest level of accuracy is 77.37%. In the other side, the lowset level of accuracy was obtained by information_gaint criterion is 72.31%. Figure 6 can be seen description the result experimtal by using DT-IG+GA and also can be seen the highest level of accuracy was obtained by parameter criterion=accuracy and gain_ratio. In this section, it is explained the results of research and at the same time is given the comprehensive discussion. Results can be presented in figures, graphs, tables and others that make the reader understand easily [2], [5]. The discussion can be made in several sub-chapters.

5. CONCLUSION

Various kinds of efforts in improving accuracy based on the results of experiments that have been done, feature selection is one of way can be used. Integrated between Information Gain and Genetic Algorithm has been give the satisfying results. Feature selection by using Gain Information and Genetic Algorithm was applied into Naive Bayes in this paper is the proposed model, that can give the best level accuracy is 92.93%. In the next study, giving the appropriate weight value to the model that will be used as learning is very influential on the level of accuracy that is produced, so there needs to be an effort in selecting the best weight. Furthur, on preprocessing process text data to this research did not do steeming process, so influence the result has been reach was not maximal. In subsequent studies need effort to increase accuracy more better than before, specially in steeming process for Indonesian text.

ACKNOWLEDGEMENTS

The authors would like to thanks the Direktorat Jenderal Penguatan Riset dan Pengembangan Kemenristek DIKTI Indonesia was funding provided through the “Penelitian Dosen Pemula (PDP)” scheme in 2018.

REFERENCES

[1] A. Muhammad, N. Wiratunga, and R. Lothian, “Contextual sentiment analysis for social media genres,” *Knowledge-Based Syst.*, vol. 108, pp. 92–101, 2016.

[2] W. Fan and M. D. Gordon, “The Power of Social Media Analytics How to use, and influence, consumer social communications to improve business performance, reputation, and profit,” *Commun. Acm*, vol. 57, no. 6, 2014.

[3] A. Reyes and P. Rosso, “Making objective decisions from subjective data: Detecting irony in customer reviews,” *Decis. Support Syst.*, vol. 53, no. 4, pp. 754–760, 2012.

[4] C. Bucur, “Using Opinion Mining Techniques in Tourism,” *Procedia Econ. Financ.*, vol. 23, no. October 2014, pp. 1666–1673, 2015.

[5] M. N. Injadat, F. Salo, and A. B. Nassif, “Data mining techniques in social media: A survey,” *Neurocomputing*, vol. 214, pp. 654–670, 2016.

[6] A. S. H. Basari, B. Hussin, I. G. P. Ananta, and J. Zeniarja, “Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization,” *Procedia Eng.*, vol. 53, pp. 453–462, 2013.

[7] A. Tripathy, A. Agrawal, and S. K. Rath, “Classification of Sentimental Reviews Using Machine Learning Techniques,” *Procedia Comput. Sci.*, vol. 57, pp. 821–829, 2015.

[8] N. Genc-Nayebi and A. Abran, “A systematic literature review: Opinion mining studies from mobile app store user reviews,” *J. Syst. Softw.*, vol. 125, pp. 207–219, 2017.

[9] D. Gräbnera and M. Zankerb, “Classification of customer reviews based on sentiment analysis.,” *Technol. Tour.* p. 12, 2012.

[10] Y. H. Hu, Y. L. Chen, and H. L. Chou, “Opinion mining from online hotel reviews – A text summarization approach,” *Inf. Process. Manag.*, vol. 53, no. 2, pp. 436–449, 2017.

[11] Y. H. Hu and K. Chen, “Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings,” *Int. J. Inf. Manage.*, vol. 36, no. 6, pp. 929–944, 2016.

[12] D. M. E. D. M. Hussein, “A survey on sentiment analysis challenges,” *J. King Saud Univ. - Eng. Sci.*, vol. 30, no. 4, pp. 330–338, 2018.

[13] K. Ravi and V. Ravi, *A survey on opinion mining and sentiment analysis: Tasks, approaches and applications*, vol. 89, no. June 2015. Elsevier B.V., 2015.

[14] B. Liu, *Sentiment Analysis and Subjectivity*, 2nd ed. Handbook of natural language processing, 2010.

[15] E. Haddi, X. Liu, and Y. Shi, “The role of text pre-processing in sentiment analysis,” *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013.

[16] M. V. Mäntylä, D. Graziotin, and M. Kuutila, “The evolution of sentiment analysis—A review of research topics, venues, and top cited papers,” *Comput. Sci. Rev.*, vol. 27, pp. 16–32, 2018.

[17] Z. E. Rasjid and R. Setiawan, “Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques,” *Procedia Comput. Sci.*, vol. 116, pp. 107–112, 2017.

[18] G. Feng, J. Guo, B.-Y. Jing, and T. Sun, “Feature subset selection using naive Bayes for text classification,” *Pattern Recognit. Lett.*, vol. 65, pp. 109–115, 2015.

[19] A. McCallum and K. Nigam, “A Comparison of Event Models for Naive Bayes Text Classification,” *AAAI/ICML-98 Work. Learn. Text Categ.*, pp. 41–48, 1998.

[20] S. P. Rajamohana and K. Umamaheswari, “Hybrid approach of improved binary particle swarm optimization and

shuffled frog leaping for feature selection," *Comput. Electr. Eng.*, vol. 67, pp. 497–508, 2018.

[21] Y. Jin, W. Xiong, and C. Wang, "Feature Selection for Chinese Text Categorization Based on Improved Particle Swarm Optimization," *Nat. Lang. Process. Knowl. Eng.*, pp. 1–6, 2010.

[22] B. M. Zahran and G. Kanaan, "Text Feature Selection using Particle Swarm Optimization Algorithm," *World Appl. Sci. JournalSpecial Issue Comput. IT*, vol. 7, pp. 69–74, 2009.

[23] Y. Liu, G. Wang, H. Chen, H. Dong, X. Zhu, and S. Wang, "An improved particle swarm optimization for feature selection," *J. Bionic Eng.*, vol. 8, no. 2, pp. 191–200, 2011.

[24] B. Xue, M. Zhang, S. Member, and W. N. Browne, "Particle Swarm Optimization for Feature Selection in Classification : A Multi-Objective Approach," pp. 1–16, 2012.

[25] J. Virmani, V. Kumar, N. Kalra, and N. Khandelwal, "SVM-based characterization of liver ultrasound images using wavelet packet texture descriptors," *J. Digit. Imaging*, vol. 26, no. 3, pp. 530–543, 2013.

[26] S. Lei, "A Feature Selection Method Based on Information Gain and Genetic Algorithm," *2012 Int. Conf. Comput. Sci. Electron. Eng.*, pp. 355–358, 2012.

[27] A. K. Uysal and S. Gunal, "Text classification using genetic algorithm oriented latent semantic features," *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5938–5947, 2014.

[28] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," *Knowledge-Based Syst.*, vol. 36, pp. 226–235, 2012.

[29] A. S. Manek, P. D. Shenoy, and M. C. M. V. K. R, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier," *World Wide Web*, 2016.

[30] T. S. Wiens, B. C. Dale, M. S. Boyce, and G. P. Kershaw, "Three way k-fold cross-validation of resource selection functions," *Ecol. Modell.*, vol. 212, no. 3–4, pp. 244–255, 2008.

[31] R. Haupt and S. Haupt, "The binary genetic algorithm," *Pract. Genet. Algorithms, Second …*, pp. 27–50, 1998.

[32] T. Weise, "Global Optimization Algorithms - Theory and Application," 2007.

[33] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Syst. Appl.*, vol. 66, pp. 1339–1351, 2016.

BIOGRAPHIES OF AUTHORS

Oman Somantri, he received his Bachelor-degree in Information Technology at the STMIK Sumedang Indonesia and later he received his Master/s Degree in Information Technology from Universitas Dian Nuswantoro Indonesia. The area of his research interest lies in data mining, sentiment analysis and Intelligent System.

Dyah Apriliani, she received his Bachelor-degree in Information Technology at the Universitas Ahmad Dahlan (UAD) Indonesia and later she received his Master/s Degree in Information System from Universitas Dipenogoro Indonesia. The area of his research interest lies in information system, and Intelligent System.