The feature extraction for classifying words on social media with the Naïve Bayes algorithm

Arif Ridho Lubis, Mahyuddin Khairuddin Matyuso Nasution, Opim Salim Sitompul, Elviawaty Muisa Zamzami
Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, Medan, Indonesia

Article Info

Article history:

Received Aug 15, 2021
Revised May 25, 2022
Accepted Jun 7, 2022

Keywords:

Feature extraction
Naïve Bayes classification
Word2Vec
Term frequency
Inverse document frequency

ABSTRACT

To classify Naïve Bayes classification (NBC), however, it is necessary to have a previous pre-processing and feature extraction. Generally, pre-processing eliminates unnecessary words while feature extraction processes these words. This paper focuses on feature extraction in which calculations and searches are used by applying word2vec while in frequency using term frequency-Inverse document frequency (TF-IDF). The process of classifying words on Twitter with 1734 tweets which are defined as a document to weight the calculation of frequency with TF-IDF with words that often come out in tweet, the value of TF-IDF decreases and vice versa. Following the achievement of the weight value of the word in the tweet, the classification is carried out using Naïve Bayes with 1734 test data, yielding an accuracy of 88.8% in the Slack word category tweet and while in the tweet category of verb 78.79%. It can be concluded that the data in the form of words available on twitter can be classified and those that refer to slack words and verbs with a fairly good level of accuracy. so that it manifests from the habit of twitter social media user.

Corresponding Author:

Mahyuddin Khairuddin Matyuso Nasution
Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara
Padang Bulan 202155 USU, Medan, Indonesia
Email: mahyuddin@usu.ac.id

1. INTRODUCTION

Based on past experience to recognize opportunities and predict the future by applying probability and statistics is the theorem of the Naïve Bayes method [1]. Strong or naive and assumptions, regardless of the condition or event of each is a character of Naïve Bayes [2]. Several data mining operations by applying Naïve Bayes with image data and numerical data from several diseases to obtain classification results [3]. In addition, classiying the behavior of web users applies naïve Bayes in the hope of obtaining optimal word segmentation results [4]. Many naïve Bayes applications classify both numerical data, images and web data with other things that are done by data crawling [5].

It is because classification is a method of using data to develop a new computational model in a certain area [6],[7]. The classification procedure employs a precise technique that differs from model to model, and a high level of accuracy is achieved when the accuracy reaches 100% [8]. It signifies that the final model produced good outcomes in terms of model creation using training and testing data. While classification by applying naïve Bayes is to detect hate speech on Twitter social media with the hope that the naïve Bayes method is able to study the previous data in Twitter and get accuracy in the carried-out test [9], by a system applying the naïve Bayes classifier by 93%. Meanwhile AlSalman [10] also conducted research on the application of naïve Bayes with the topic of sentiment analysis on social media content to get opinions from several different applications and fields such as hobbies, activities and work carried out in Twitter which uses

Arabic, the results of the experiment get useful from the proposed approach proposed to be continued. From these results were obtained comparisons showing this approach outperforms the field of work and it can increase the accuracy of 0.3%. Of various studies, naïve Bayes is often used in classification on social media to get sentiment analysis [11]. At this time, social media is one aspect that is very close to users, social media users are used to creating and sharing content any between users [12]-[15]. Runining social media about 142 minutes a day [16], the initial low increases 100 minutes to get 142 minutes per day of use [17]. It is difficult to identify whether such platforms are profitable or detrimental to social media users, even though people around the world spend a large part of their days on social media platforms.

This is related to the research conducted by Lubis *et al.* [18] finding a framework for social media users, both the disclosure of words on social media is the keyword as the habit of social media users with initialized steps by reviewing current postings in order to have good data that is more particular and precise than those obtained from netizens [19]. These exact keywords can then be used by the social media system to detect the profile of a certain user in the online domain in which a search engine can access. This certainly opens up insights that the behavior of social media users could be classified by applying the naïve Bayes method with training data in the form of words and keywords to classify words on social media [20]. In matching and obtaining the frequency of the word data, nevertheless, a feature extraction stage is needed. So that in this study a comparison of several feature extraction techniques was carried out to obtain the optimal classification process.

So many feature extractions are available that the research focuses on feature extraction on word classification on social media using naïve Bayes. In line with the development of data science applied in this paper, however, feature extraction uses term frequency-Inverse document frequency (TF-IDF) and Word2Vec. Where Word2Vec predicts the word given the surrounding context and after the occurrence of the model is created, what context vector operation is appropriate to perform the task for classification on the word in the new tweet [21].

2. MATERIAL AND METHOD

2.1. Data Mining in Social Network

Data mining is a term that usually refers to knowledge findings in databases. It is a process that practices mathematical, statistical, artificial intelligence, and machine education methods that extract and recognize useful data and knowledge gathered from large databases [22]. Data mining, furthermore, is also referred to as the process of finding patterns, trends, and meaningful relationships [23]. Before carrying out the data mining process, it is better to know in advance what data mining can do, so that what is done later is suitable with what is needed and produces something that was previously unknown and is new and useful for its own users [24]. In principle, data mining has several tasks and must ensure that the pattern runs correctly in the process. There are 2 types of information mining tasks, namely [25]:

a. Predictive

   Estimating a certain attribute's value based on the values of other attributes. The dependent and target variables in such a case are called attributes, while the independent variable attribute is used to predict

b. Descriptive

   Obtaining patterns such as groups, trajectories, correlations, anomalies, and trends, which summarize the underlying relationships in the data is the task of descriptive. Descriptive data mining tasks are also known as investigations and often require post-processing techniques for explanation and validation of the results.

2.2. Naïve Bayes Algorithm

One of the methods of classification is Naïve Bayes, this algorithm was invented by Thomas Bayes who is a scientist from England. Future opportunities can be predicted based on previous experience is the goal of Naïve Bayes [26]. This Naïve Bayes Classifier has the main characteristic of being very strong (naive) assumptions about each condition's self-sufficiency. In compared to other classifier models, the Naïve Bayes Classifier performs quite well. One of the benefits of this method is that it just takes a little amount of training data to calculate the parameter estimation used in classification. The independent variable is the variation of a variable in a class that is designed to decide classification, not the whole covariance matrix [27].

The training stage and the classification stage are the stages of Naïve Bayes. The process of analyzing the document is carried out at the training stage where the vocabulary selection of the sample document is the word that appears in the sample document where the word is a representation of the document. The next step is to determine the probability for each category based on a sample document. Naïve Bayes built a probabilistic model from the term documents matrix data labeled. Document classification is done by first determining the category c words in the document. The process of determining the categories of a document is done by calculating using equation (1) [28]:

$$c^8 = argmax_{ci\in c} p(c_i|d_j) = argmax_{ci\in c} \prod p\,(w_{kj}|c_i)xp(c_i) \tag{1}$$

where:
w_{kj} is a feature or word of the document / tweet
d_j category to find out
The value of p (wkj | ci) is known from the available training data.

3. GENERAL ARCHITECTURE

Good research requires a research flow. The purpose of the research flow is to describe the stages that are carried out, where these stages are well explained. The research flow itself is used to ensure the research runs as expected. The flow of the research is drawn in a general structure as depicted in Figure 1:

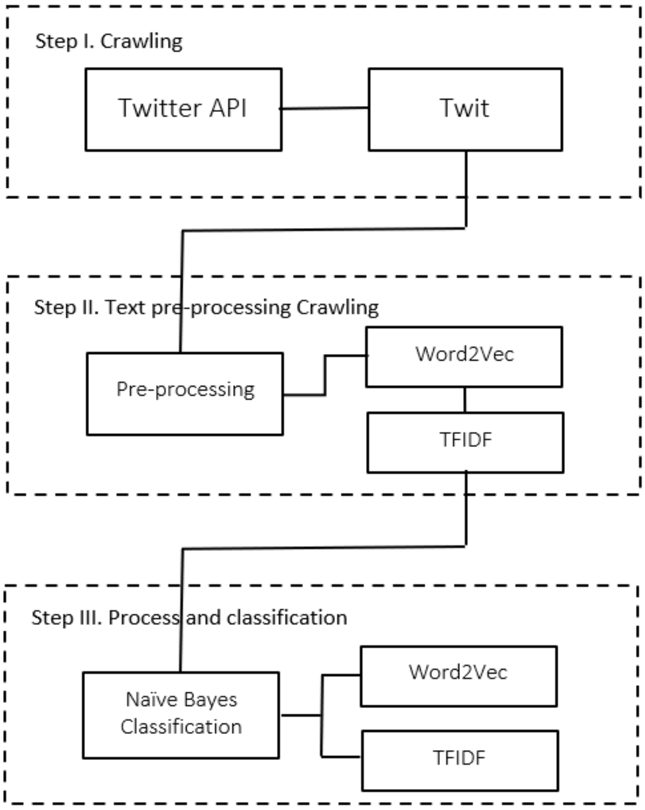
Step I. Crawling
Twitter API
Twit
Step II. Text pre-processing Crawling
Pre-processing
Word2Vec
TFIDF
Step III. Process and classification
Naïve Bayes Classification
Word2Vec
TFIDF

Figure 1. General Architecture

The explanation of the general architecture in Figure 1 is:

a) Step 1. The crawling process using the API on Twitter makes it easy to get tweets that will be classified.
b) Step 2. The text preprocessing process is then continued with the feature extraction process where the feature extraction process is optimized by counting words assisted by Word2Vec then calculating the frequency with the IDF TF as well as contributing to this paper.
c) Step 3. The classification process applies NBC with the results, word classification and accuracy. The NBC procedure consists of the following steps:
i) Making decomposition data
ii) Reading the training data
iii) For numeric data, how to calculate the number and probability is
– Each parameter is numeric, then the mean and standard deviation are calculated. The (2) to find the calculated average (mean) is as in (2):

$$\mu = \frac{\sum_i^n x_i}{n} \tag{2}$$

where
μ : mean
x_i : the value of x to i
n : Total samples
While the (3) is to find the score/ value, Deviation Standard can be:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n-1}} \tag{3}$$

where
σ : deviation standard
x_i : the value of x to i
μ : mean
n : Total samples

– To get a probabilistic value, divide the amount of acceptable data from the same category by the amount of data in that category that is included.

iv) Get value in word classification
v) Produce accuracy with (4)

$$Accuracy = \frac{Number\ of\ correct\ classifications}{Amount\ of\ test\ data} X100\% \tag{4}$$

4. RESULT AND DISCUSSION

The data mining process in the form of classification techniques can be done using the Naïve Bayes Algorithm. Naïve Bayes generally are also often used in research that is sentiment analysis to get accuracy, patterns, human behavior and others available in cloud networks. The classification process with naïve Bayes cannot be separated from the process of training and testing data so that the correct data in this study use and collect data taken from social media. The data crawling is then performed feature extraction to facilitate the classification of words on social media. Several feature extractions will be tested to optimize the word classification using naïve Bayes on social media.

Terms, which can be a sentence, word, or other indexing unit in a tweet that serves to establish the context, are things to consider while looking for information from a collection of documents or tweets. Because each word has a distinct amount of relevance in the tweets on the tab, an indicator, specifically the term weight, is supplied for each word. When using word2vec to count and search for words, there are a few things to keep in mind. The results are shown in Figure 2:

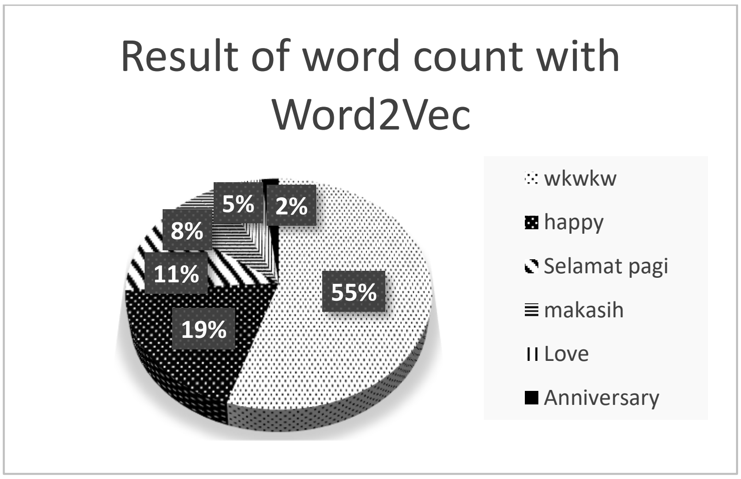

Figure 2. Results of Processing with Word2Vec

Preprocessing and feature extraction are the next steps in the classification process, which will be used to find meaning in tweets that will be trained or tested. This procedure must be followed since the document test data is in the form of paragraphs containing labels obscuring its content. Before the preparation process, it was difficult to understand the contents of the test text. Features that can potentially be affected by preprocessing so it is necessary to identify the text.

Tokenization is first performed which aims to separate characters into tokens or words. As certain characters could be used to separate tokens, tokenization is difficult for computer programs. To detect the pattern of the text which is going to be used for the categories that will be used as training data, so text identification is carried out.

Next perform frequency calculations with TF-IDF. Where, the term weighting method which is commonly used as a comparison method with the new weighting method is TF-IDF. T term weight calculation of a document is done by multiplying the value of the Term Frequency Inverse Document Frequency. Some of the processes taken to compute the weight value using TF-IDF.

Table 1 reveals that TF-IDF calculations were performed using the frequency of tweets that frequently appear, namely the term "wkwk," which is slang for anyone who laughs or receives amusing things, containing the word happy, which is a verb in the form of joy. Where, the TF-IDF calculation requires a frequency-weighted value that has a function to calculate the best value, this is because the higher the number of words when calculating the TF-IDF, the smaller the frequency. To find terms and assess the performance of documents which are based on tweets that appear, NBC is used. calculating the frequency of occurrence of words in the document is the first step taken. where the high frequency of repetition causes the greater the value of the word.

Table 1. Terms of Optimization Word2Vec and TF IDF

Word (t)	TF	IDF
happy	39	=log(112/39)= 0.4578818967
makasih	17	=log(112/17)= 0.8188854146
Anniversary	4	=log(112/4)= 1.447158031
wkwkwk	112	=log(112/112)=0
Love	9	=log(112/9)= 1.09482038
Selamat pagi	22	=log(112/22)= 0.7067177823

The NBC method requires two stages in the word classification process, namely the first stage is training where at this stage analysis are conducted on the documents of sample. They are in the format of social media data, namely tweets, words which might be shown up within a collection of documents of sample as well as determined from people's habits on social media reflect as many documents as possible, the documents used for training will be a reference in the testing process, as shown in Table 2. In the second stage is testing there is a training document that will be used as a reference for the testing process.

Table 2. Data Decomposition

Word (t)	Training	Testing
happy	75%	25%
makasih	75%	25%
Anniversary	75%	25%
wkwkwk	75%	25%
Love	75%	25%
Selamat pagi	75%	25%

In this study, data sources were employed from Twitter classified into documents as a reference to how papers would be classed. The targeted reference is document labeling based on expert domains. Twitter social media tweets are the type of document used. Twitter itself is unstructured content because there are things like mentions and HTML tags that cause the document to be meaningless. For classification accuracy, a structured document is needed so that it is easy to understand. The experimental document consists of 1734 characters from the @arfridho account.

In the previous stage, Despite the fact that the generated text pattern was analysed by applying stopwords, the irregularity of the produced text pattern presents a challenge during identification. It appears that identifying the text was difficult and that careful examination was necessary. Because the patterns are irregular in their content arrangement, it is necessary to read the documents one at a time throughout the identification process to grasp the existing patterns in the text. The procedure to identify the tag on the training document is manually conducted. The data will be divided into two groups: slack words and verbs. Term frequency can be used to determine characteristics; however, experimental data indicated that only 25% of the chosen terms occurred frequently, which has little bearing on the categorization process.

The data in Figure 3 are those having been achieved using the IDF TF. Then the document classification process requires a calculation involving the number of label documents n, m is the number of label documents, and the total number of training documents, which is called p (ci), namely the category x dividing the total documents in the category x the number of training data, similar to the category y is the division of the number of categorized documents y with the total number of training data, as shown in Table 3. In research with data with words on Twitter that were tested as many as 531 twitting of verb categories and 1734 tweets referring to 1203 categories of Slack words as training data resulted in accuracy calculated based on (4) well in the tweet category of verbs 78.79% whereas on the Slack word category tweet of 88.8%.

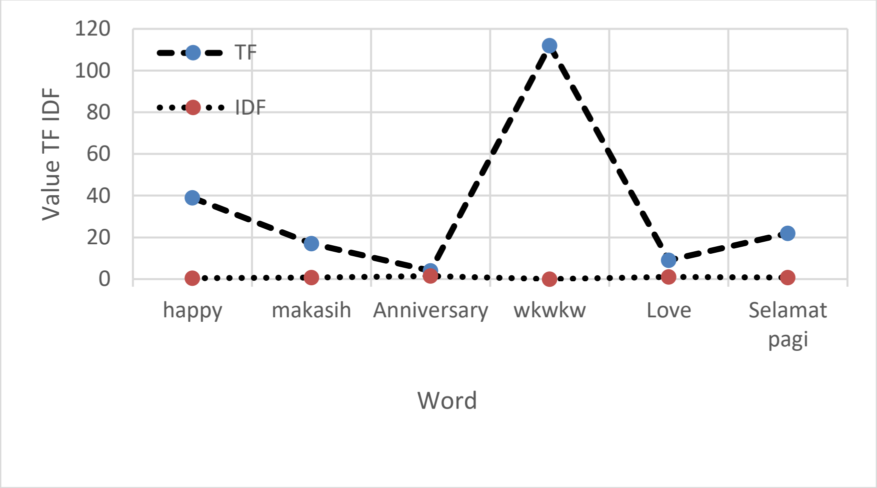

Figure 3. Classification features

Table 3. Social Media Word Classifications by NBC

Category	Word Classification	Verb	Slack words
p(c$_i$)		0.50	0.50
P(w$_{kj}$|c$_i$)	happy	15	24
	makasih	17	0
	Anniversary	2	2
	wkwkw	0	112
	love	9	0
	Selamat pagi	20	2

5. CONCLUSION

This paper draws the conclusion that data in the form of words available on twitter can be classified and those refering to the word slack and verb so as to manifest from the habit of the social media twitter users. In the process, word classifications in social media are conducted, beginning with data crawling on the Twitter API then carrying out preprocessing and feature extraction. Where there is interest in the feature extraction process with a combination of word2vec with TF-IDF where the results of the TF-IDF calculation result is possible to deduce that the TF-IDF value with frequencies which frequently appear gets smaller frequency values and conversely with less frequency than the value of TF -IDF is even bigger. Following the TF-IDF calculation, the classification is done using the Naïve Bayes approach, which divides the word classification into two categories, namely the Slack word category and the verb category. The test data consisted of 1734 twittes, with the results referring to 1203 Slack word categories and 531 twittens of verb categories as training data, resulting in good accuracy in the Slack word category twitt of 88.8% and 78.79% in the twitt verb categories. Where the results obtained from the test, obtained a fairly good accuracy in categorizing slack words and verbs.

REFERENCES

[1] E. Sugiharti, S. Firmansyah, and F. R. Devi, "Predictive evaluation of performance of computer science students of unnes using data mining based on naÏve bayes classifier (NBC) algorithm," *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 4, pp. 902–911, 2017.

[2] H. Zhang, "Exploring Conditions for the Optimality of Naïve Bayes," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 19, no. 02, pp. 183–198, Mar. 2005, doi: 10.1142/S0218001405003983.

[3] S. B. Özkan, S. M. F. Apaydin, Y. Özkan, and I. Düzdar, "Comparison of Open Source Data Mining Tools: Naive Bayes Algorithm Example," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 2019, pp. 1–4, doi: 10.1109/EBBT.2019.8741664.

[4] D. Bai, L. Zeng, and M. Feng, "Naive bayes for web penetration behavior classification based on word segmentation improvement," in *2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, 2019, pp. 419–422, doi: 10.1109/ICISCAE48440.2019.221666.

[5] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," *Expert Syst. Appl.*, vol. 36, no. 3, Part 2, pp. 6527–6535, 2009, doi: https://doi.org/10.1016/j.eswa.2008.07.035.

[6] M. Morgan, C. Blank, and R. Seetan, "Plant disease prediction using classification algorithms," *IAES Int. J. Artif. Intell.*, vol. 10, no. 1, pp. 257–264, 2021, doi: 10.11591/ijai.v10.i1.pp257-264.

[7] A. R. Lubis, M. Lubis, and Al-Khowarizmi, "Optimization of distance formula in k-nearest neighbor method," *Bull. Electr. Eng. Informatics*, vol. 9, no. 1, pp. 326–338, 2020, doi: 10.11591/eei.v9i1.1464.

[8] A. R. Lubis, M. Lubis, Al-Khowarizmi, and D. Listriani, "Big Data Forecasting Applied Nearest Neighbor Method," in *ICSECC 2019 - International Conference on Sustainable Engineering and Creative Computing: New Idea, New Innovation, Proceedings*, 2019, pp. 116–120, doi: 10.1109/ICSECC.2019.8907010.

[9] N. R. Fatahillah, P. Suryati, and C. Haryawan, "Implementation of Naive Bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech," in *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*, 2017, pp. 128–131, doi: 10.1109/SIET.2017.8304122.

[10] H. AlSalman, "An Improved Approach for Sentiment Analysis of Arabic Tweets in Twitter Social Media," in *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*, 2020, pp. 1–4, doi: 10.1109/ICCAIS48893.2020.9096850.

[11] C. Fiarni, H. Maharani, and R. Pratama, "Sentiment analysis system for Indonesia online retail shop review using hierarchy Naive Bayes technique," in *2016 4th International Conference on Information and Communication Technology (ICoICT)*, 2016, pp. 1–6, doi: 10.1109/ICoICT.2016.7571912.

[12] K. Subrahmanyam, S. M. Reich, N. Waechter, and G. Espinoza, "Online and offline social networks: Use of social networking sites by emerging adults," *J. Appl. Dev. Psychol.*, vol. 29, no. 6, pp. 420–433, 2008, doi: https://doi.org/10.1016/j.appdev.2008.07.003.

[13] N. S. Shaeeali, A. Mohamed, and S. Mutalib, "Customer reviews analytics on food delivery services in social media: A review," *IAES Int. J. Artif. Intell.*, vol. 9, no. 4, pp. 691–699, 2020, doi: 10.11591/ijai.v9.i4.pp691-699.

[14] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Measuring information credibility in social media using combination of user profile and message content dimensions," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 4, pp. 3537–3549, 2020, doi: 10.11591/ijece.v10i4.pp3537-3549.

[15] N. S. A. Rahman, L. Handayani, M. S. Othman, W. M. Al-Rahmi, S. Kasim, and T. Sutikno, "Social media for collaborative learning," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 1, pp. 1070–1078, 2020, doi: 10.11591/ijece.v10i1.pp1070-1078.

[16] T. Roshini, P. V Sireesha, D. Parasa, and S. Bano, "Social Media Survey using Decision Tree and Naive Bayes Classification," in *2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT)*, 2019, pp. 265–270, doi: 10.1109/ICCT46177.2019.8969058.

[17] E. J. Ivie, A. Pettitt, L. J. Moses, and N. B. Allen, "A meta-analysis of the association between adolescent social media use and depressive symptoms," *J. Affect. Disord.*, vol. 275, pp. 165–174, 2020, doi: https://doi.org/10.1016/j.jad.2020.06.014.

[18] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, "A Framework of Utilizing Big Data of Social Media to Find Out the Habits of Users Using Keyword," 2020, pp. 140–144.

[19] A. R. Lubis *et al.*, "Obtaining Value From The Constraints in Finding User Habitual Words," pp. 8–11, 2020.

[20] M. Asif, A. Ishtiaq, H. Ahmad, H. Aljuaid, and J. Shah, "Sentiment analysis of extremism in social media from textual information," *Telemat. Informatics*, vol. 48, p. 101345, 2020, doi: https://doi.org/10.1016/j.tele.2020.101345.

[21] S. Lei, "Research on the Improved Word2Vec Optimization Strategy Based on Statistical Language Model," in *2020 International Conference on Information Science, Parallel and Distributed Systems (ISPDS)*, 2020, pp. 356–359, doi: 10.1109/ISPDS51347.2020.00082.

[22] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," *Expert Syst. Appl.*, vol. 166, p. 114060, 2021, doi: https://doi.org/10.1016/j.eswa.2020.114060.

[23] S. Shirowzhan, S. Lim, J. Trinder, H. Li, and S. M. E. Sepasgozar, "Data mining for recognition of spatial distribution patterns of building heights using airborne lidar data," *Adv. Eng. Informatics*, vol. 43, p. 101033, 2020, doi: https://doi.org/10.1016/j.aei.2020.101033.

[24] H. Thakkar, V. Shah, H. Yagnik, and M. Shah, "Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis," *Clin. eHealth*, vol. 4, pp. 12–23, 2021, doi: 10.1016/j.ceh.2020.11.001.

[25] J. Han, M. Kamber, and J. Pei, "1 - Introduction," in *The Morgan Kaufmann Series in Data Management Systems*, J. Han, M. Kamber, and J. B. T.-D. M. (Third E. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 1–38.

[26] P. Phoenix, R. Sudaryono, and D. Suhartono, "Classifying Promotion Images Using Optical Character Recognition and Naïve Bayes Classifier," *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 498–506, 2021, doi: 10.1016/j.procs.2021.01.033.

[27] S. Theodoridis, "Chapter 12 - Bayesian Learning: Inference and the EM Algorithm," S. B. T.-M. L. (Second E. Theodoridis, Ed. Academic Press, 2020, pp. 595–646.

[28] S. Theodoridis, "Chapter 13 - Bayesian Learning: Approximate Inference and Nonparametric Models," S. B. T.-M. L. (Second E. Theodoridis, Ed. Academic Press, 2020, pp. 647–730.

BIOGRAPHIES OF AUTHORS

Arif Ridho Lubis He got master from Universiti Utara Malaysia in 2012 and graduate from Universiti Utara Malaysia in 2011, both information technology. He is a lecturer in Department of Computer Engineering and Informatics, Politeknik Negeri Medan in 2015. His research interest includes computer science, network, science and project management. He can be contacted at email: arifridho.l@students.usu.ac.id.

Mahyuddin Khairuddin Matyuso Nasution Professor from Universitas Sumatera Utara, Medan Indonesia. Mahyuddin K. M. Nasution was born in the village of Teluk Pulai Dalam, Labuhan Batu Regency, North Sumatera Province. Worked as a Lecturer at the Universitas Sumatera Utara, fields: Mathematics, Computer and Information Technology. Education: Drs. Mathematics (USU Medan, 1992); MIT, Computers and Information Technology (UKM Malaysia, 2003); Ph.D. in Information Science (Malaysian UKM). He can be contacted at email: mahyuddin@usu.ac.id.

Opim Salim Sitompul Received the Ph.D. degree in information science from Universiti Kebangsaan Malaysia, Selangor, in 2005. He is currently a Professor with the Department of Information Technology, Universitas Sumatera Utara, Medan, Indonesia. His skill and expertise are in AI, data warehousing, and data science. His recent projects are in natural language generation (NLP) and AIoT. The result of his work can be seen in his most recent publication is ‘‘Template-Based Natural Language Generation in Interpreting Laboratory Blood Test.’ He can be contacted at email: opim@usu.ac.id.

Elviawaty Muisa Zamzami Graduated from Bandung Institute of Technology (Indonesia), magister of informatics, 2000 and awarded Doctoral in Computer Science from the University of Indonesia in 2013. She is a lecturer at Department of Computer Science, Universitas Sumatera Utara, Indonesia. Currently her research interests are reverse engineering, requirements recovery, software engineering, requirements engineering and ontology. She can be contacted at email: elvi_zamzami@usu.ac.id.