International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Sentiment Analysis on E-commerce Product using Machine Learning and Combination of TF-IDF and Backward Elimination

Tommy Willianto, Supryadi, Antoni Wibowo

Abstract: ***E-commerce is a website or mobile application platform that help people to buy products. Before purchasing the product, customer will decide to buy it or not by reading the review from previous buyer. There is a problem that there are a lot of review so it will take a long time for customer to read it all. This research will be using sentiment analysis method to classify the review data. Sentiment analysis or opinion mining is a machine learning approach to classify and analyse texts or documents about human's sentiments, emotions, and opinions. In this research, sentiment analysis was used to classify product reviews from e-commerce websites into positive or negative classes. The results could be processed further and be used to summarize customers' opinions about a certain product without reading every single review. The goal of this research is to optimize classification performance by using feature selection technique. Terms Frequency-Inverse Document Frequency (TF-IDF) feature extraction, Backward Elimination feature selection, and five different classifiers (Naïve Bayes, Support Vector Machine, K-Nearest Neighbour, Decision Tree, Random Forest) were used in analysing the sentiment of the reviews. In this research, the dataset used are Indonesian language and classified into two classes(positive and negative). The best accuracy is achieved by using TF-IDF, Backward Elimination and Support Vector Machine (SVM) with a score of 85.97%, which increases by 7.91% if compared to the process without feature selection. Based on the results, Backward Elimination feature selection succeeded in improving all performance for all classifiers used in this research.***

Keywords : Backward elimination, e-commerce, sentiment-analysis, TF-IDF (terms frequency-inverse document frequency)

I. INTRODUCTION

In this era, information technology plays a big role to make people's life more comfortable. Even purchasing and selling products become easier with online e-commerce, so people don't need to go outside to buy things. E-commerce (electronic commerce) is selling, distributing, and marketing products and services using electronic systems [1]. Tokopedia, Shopee, and Bukalapak are the three best e-commerce in Indonesia currently [2].

In marketplace's web application and mobile application, there are features of reviews and ratings that can be given by customers after making a product purchase. Many e-commerce websites encouraging online users to post their evaluations on product or the service. Reviews and ratings can be information for sellers to find out feedback from customers, also useful for other customers to find out whether the store is reliable and good [3]. However, if there are many purchase transactions, there are also lots of reviews on the product. Too many reviews make users lazy to read everything one by one.

Sentiment analysis or opinion mining is a machine learning approach to classify and analyse texts or documents about human's sentiments, emotions, opinions. It will classify texts into some classes according to the amount of label from data training. With machine learning, reading all of the reviews is rather time consuming where we can summarize the review on particular category [4]. The importance of sentiment analysis is increasing as the amount of opinion data increases. So the machine needs to be more reliable and efficient [5].

In this paper, the research focuses on combining TF-IDF and feature selection with different classification algorithms. The algorithms that are used in this research are Support Vector Machine (SVM), Naive Bayes, Decision Tree, K-Nearest Neighbour (K-NN), and Random Forest. The feature selection using forward selection algorithm.

II. RELATED WORKS

Kamilah research about sentiment analysis for product review using Naive Bayes algorithm. The dataset retrieved from the Tokopedia website and translated to English, with 200 dataset training and 20 dataset testing. Data pre-processing with tokenization, filtering, stemming, and transformation. Then the data classified to positive and negative classes using Naive Bayes algorithm. Validation with cross validation. Accuracy result 77% [6].

Different algorithms are used in Hario research. The research also compares extraction methods between Terms Frequency-Inverse Document Frequency (TF-IDF) and N-gram using SVM classifier. The retrieved datasets are Bahasa Indonesia.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Tommy Willianto, Computer Science Department, Binus Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480. Email: tommy.willianto@binus.ac.id
Supryadi, Computer Science Department, Binus Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480. Email: supryadi@binus.ac.id
Antoni Wibowo, Computer Science Department, Binus Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480. Email: anwibowo@binus.ac.id

The best result obtained from using unigram and SVM method with 80,87% accuracy [7].

Billy works on sentiment analysis using Naive Bayes algorithm with different amounts of classes and data training. The datasets are Bahasa Indonesia sentiments classified into 3 classes (positif, netral, negatif) and 5 classes (sangat positif, positif, netral, negatif, sangat negatif). The dataset split into different amounts of data training, 80% and 90%[6]. The highest accuracy achieved was 77.78% that classified into 3 classes and using 90% data training [8].

Twitter users' opinions about the service of the marketplaces are used in Muljono work. The dataset retrieved with crawling opinions from twitter using twitter API. Collected data was 1200 Bahasa Indonesia opinion data [9]. After pre-processing, the data weighted using TF-IDF. The data classified using Naive Bayes algorithm. Accuracy achieved was 93.3%.

Sudheer also works on real time sentiment analysis of tweets about e-commerce websites data. The collected tweets about e-commerce are Amazon 50000 tweets, eBay 25000 tweets, and Alibaba 25000 tweets. The work focuses on comparing accuracy with different classifiers, feature selection, and datasets. The algorithms used in this paper are Naive Bayes, Maximum Entropy, and Decision Tree. The feature selection used are document frequency and part of speech tag. The finest result is data set from amazon e-commerce, much of the time Naive Bayes classifier outperformed the other classifier [10].

Table-I: Related Works

No	Title	Author	Problem	Method	Result
1	Analisa Sentimen Pelanggan Tokopedia Menggunakan Algoritma Naive Bayes Berdasarkan Review Pelanggan	Ai Nurhayatul Kamilah	Product quality has not been conveyed properly to customers	Using Naive Bayes Classification, Tokopedia product review dataset translated to English, 200 data training, 20 data testing	Accuracy 77%
2	Support Vector Machine Classifier for Sentiment Analysis of Feedback Marketplace with a Comparison Features at Aspect Level	Hario Laskito Ardi, Eko Sediyono, Retno Kusumaningrum	Compare characteristics analysis to get the best classification results.	TF-IDF, n-gram using Support Vector Machine	The unigram character analysis model and the Support Vector Machine classification are the best models with an accuracy value of 80.87%.
3	Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes	Billy Gunawan, Helen Sasty Pratiwi, Enda Esyudha Pratama	Difficulty to read all of the reviews and opinions because the data too much	Classification with Naive Bayes Algorithm. 4 testing, classified into 3 classes and 5 classes with 2 different amount data training	Accuracy 77.78% (3 class and 90% data training), 73.89%(3 class and 80% data training), 59.33%(5 class and 90% data training), 52.66%(5 class and 80% data training)
4	Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naive Bayes	Muljono, Dian Putri Artanti, Abdul Syukur, Adi Prihandono, De Rosal I. Moses Setiadi	Consumer use social media to express their opinion about the services of online marketplace	Using Naive Bayes Classification Algorithm, 1200 dataset from twitter.	Accuracy 93.3%
5	Real Time Sentiment Analysis of E-Commerce Websites Using Machine Learning Algorithms	Prof. K. Sudheer, Dr. B. Valarmathi	The e-commerce websites only maintain positive rating	50000 datasets from amazon tweets, 25000 dataset from eBay tweets, 25000 dataset from Alibaba tweets	The best accuracy 92% obtained from amazon using document frequency feature selection

III. METHODOLOGY

Sentiment analysis is a computational methodology to identify and extract the sentiment contents in text, speech, or database. Sentiment analysis also characterized emotions, subjective impression, and opinions [11]. Classifying sentiments on e-commerce product review with the best performance is the main goal in this paper. Sentiments will be classified into either positive or negative classes. Fig. 1. shows the steps of the proposed work to accomplish the expected results.

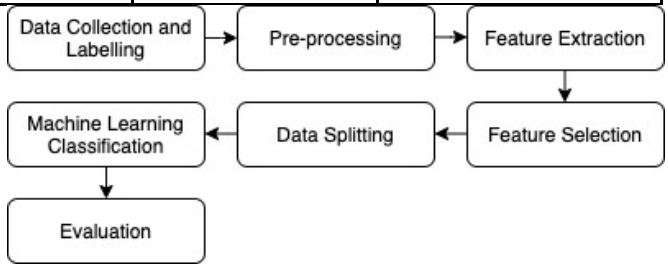

Fig. 1. Research Methodology

A. Data Collection and Labelling

The first step in this research methodology is data collection from the marketplace's website using web scraping. Mitchell(2018) defined that web scraping is gathering data sourced from the internet. It works with accessing the web page, selecting the data elements, extract and store it into a structured dataset [12].

Data is collected by scraping from web pages with the help of Data Miner Google Chrome Extension. A total of 1500 reviews is acquired equally from the official website of Tokopedia, Shopee, and Bukalapak in order to obtain a variety of data. The collected data will then be exported to an Excel spreadsheet. After data is successfully collected, the labelling process will take place. Labelling is a process of tagging wherein in this case is to set a positive or negative label to classify the sentiment of a review. Results of this process is a dataset that contains two classes of sentiments and is ready to be processed. Below are some examples of dataset that have been labelled.

Review	Sentiment
Barang sesuai pesanan dan cepat sampai	Positive
Pengiriman cepat. Fast respon. Barang sesuai. Trims	Positive
Barang komplit dah tidak ada kekurangan, terima kasih kak 😊	Positive
Bintang 1 karna Pop Socket tidak dikirim 😥	Negative
Respon lama, seller tdk komunikatif	Negative

Fig. 2. Labeled Dataset Examples

B. Pre-processing

Data that is collected from the previous step must undergo several cleansing operations in order for it to be processed and used in machine learning. This step is also known as pre-processing. Pre-processing is one of the most important steps in data mining. A good result definitely depends on how well the data is handled. There are a lot of pre-processing techniques. There are filtering, stemming, and tokenizing [13]. Pre-processing in this paper is done with the help of RapidMiner software. Below are the steps that will be done to filter the data.

- Remove Duplicates and Missing Values
  Often in large amounts of data, there will exist some duplicate and missing values. These values are called noise and could interfere with the performance of the model. Therefore it is important to remove duplicate and missing values in the data.
- Replace Emoji and Emoticon
  Emoji and Emoticon are useful in expressing feelings in a sentiment. It is a very strong way to represent human's feeling. Emoji and Emoticon could be used alone or even with words to clarify the meaning of a sentence. Since sentiment analysis uses text mining, emoji and emoticons should be converted into text in order for the machine to understand [14]. Below are some examples of emoji and emoticons that will be converted to text.

:)	Positive
:(	Negative
😠	Negative
❤️	Positive
😍	Positive

Fig. 3. Examples of Emoji and Emoticons that will be Converted

- Stemming
  The sentiment analysis will work better with stem words, so the words need to be transformed to the original form. Stemming is transforming the word into its stem form by removing prefix, suffix, and infix. In this research, the stemming process using PHP library from https://github.com/sastrawi/sastrawi [15].
- Transform Cases
  A word consists of letters with different cases such as upper case and lower case. To standardize letter cases, all letters are converted to its lower case. Transforming cases also improves the consistency of the data. [16]
- Filter Stopwords
  Stopword is a vocabulary that is not unique words from a document. The examples are "di", "oleh", "pada", "sebuah", "karena", etc. The stopwords will be removed to improve the performance of sentiment analysis [17].
- Tokenize and Filter Tokens by Length
  Tokenize is a process for separating words and resulting tokens. Then the generated token filtered according to the length of the character [18].

C. Feature Extraction

Feature extraction is the process of dimensionality reduction which transforms original data to a dataset with reduced number of variables. These variables are also known as features. Feature extraction is effective in reducing the amount of data that needs to be processed while still maintaining relevant information of the original dataset. Feature extraction can also reduce redundant data in a dataset and speeds up machine learning process [19].

- TF-IDF
  Term frequency inverse document frequency (TF-IDF) is a method to calculate the weight value of word(term) contained in a document. Term-frequency measuring the frequency of a term appears in a document(1). While the inverse document frequency is logarithm of the ratio of the total number of documents in the number of corpus by the number of documents that have the term(2). The equation of TF-IDF is represented by the following equation(3) [20]:

$$tf_i = \frac{freq_i(d_j)}{\sum_{i=1}^{k} freq_i(d_j)} \quad (1)$$

$$idf_i = log \frac{|D|}{|\{d:t_i \in d\}|} \quad (2)$$

$$(tf - idf)_{ij} = tf_i(d_j).idf_i \quad (3)$$

D. Feature Selection

Feature selection is one of the main data mining tasks. It helps in selecting the most relevant features for classification. The irrelevant and redundant features may confusing the classifier and lead to incorrect results. The use of feature selection will reduce the dimensionality of dataset and increase the learning accuracy [21].

- Backward Elimination
  There are a lot of feature selection methods. In this research, backward elimination is used to improve the process by selecting the most relevant attributes and eliminating rare and unused features. The Backward Elimination feature selection starts with the full set of attributes and in each iteration, it removes each remaining attribute of the given dataset. For each removed attribute, the performance is estimated using the inner operators. The attribute that gives the least decrease of performance will be removed from the selection. Then a new iteration is started with the modified selection. This implementation avoids any additional memory consumption besides the memory used originally for storing the data and the memory which might be needed for applying the inner operators [22].

E. Data Splitting

In order to continue the process, data needs to be split into two parts. The first part is for training and the other part is for testing purposes. The ratio of splitting this dataset is 80% for training and 20% for testing. Training data will be used to train the machine in order to classify sentiments of the reviews. While testing data is used to test the performance of the model that has been trained.

F. Machine Learning Classification

Text classification has been studied in different communities of information technology, such as data mining, database, machine learning, and information retrieval. The goal of text classification is to assign predefined classes to text documents. There are many applications of text classification, such as image processing, medical diagnosis, document or organization, etc [23]. There are many algorithms for classifying data where each of it produce different performance results. In this study, 5 algorithms are used in order to achieve the best results.

- Support Vector Machine (SVM)
  SVM is a classifier that is defined by a separating hyperplane. The goal of SVM is to find the optimal separating hyper-plane (OSH) that has the maximal margin to both sides [24].
- Naive Bayes
  Naive Bayes is an algorithm to find the highest probability value to classify the data testing into the most proper category. A very strong assumption of independence from each condition or event is the main characteristic of Naive Bayes. Each document is represented with a pair of attributes "x1, x2, ...xn" where x1 is the first word, x2 is the second word, etc [25].
- Decision Tree
  Decision trees are considered as one of the most popular data-mining techniques. Decision tree splits recursively a dataset of records using depth-first approach or breadth-first approach. The process works until all data items have been classified. This algorithm is desirable for small-medium data sets [26].
- K-Nearest Neighbour
  KNN is a vector space model method for classifying objects. KNN classifies the object with theory a test document d will have the same category or label as the category of the training document positioned in the scope of k surrounding the document d. The parameter k in KNN is often chosen based on experience or knowledge of the classification problem at hand [27].
- Random Forest
  Random forest is a supervised machine learning algorithm based on ensemble learning(forming a powerful prediction model with joining different or the same algorithm multiple times). So this algorithm combines multiple decision tree algorithms and resulting a forest of trees. Random forest splits each node in tree using the best among subset of predictors randomly chosen at the node [28].

G. Evaluation

The confusion matrix evaluation is used after classifying the data. The table of confusion matrix is shown in Fig. 4. True Positive (TP) is the number of data that labelled as positive and classified as positive by classifier. False Positive (FP) is the number of data that labelled as positive, but classified as negative by classifier. True Negative (TN) is the number of data that labelled as negative and classified as negative by classifier. False Negative (FN) is the number of data that labelled as negative, but classified as positive by classifier.

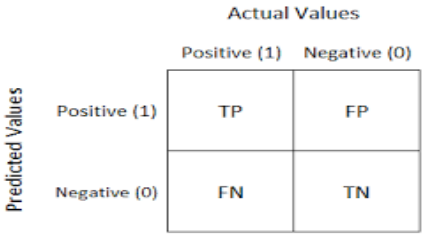

Fig. 4. Confusion Matrix Table [29]

The performance of the classifier is measured with accuracy, precision, recall, and f-measure [30].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F-measure = \frac{2\times Recall\times Precision}{(Recall+Precision)} \quad (7)$$

IV. RESULTS AND DISCUSSION

The focus of this paper is to prove that feature selection can be an option to improve performance accuracy in sentiment analysis. The feature selection that is used in this research is Backward Elimination. TF-IDF and Backward Elimination are combined and used in the following classification operators: SVM, Naive Bayes, Decision Tree, K-NN, and Random Forest.

Table-II: Confusion Matrix

Method	TF-IDF				TF-IDF & BE			
	TP	FP	TN	FN	TP	FP	TN	FN
SVM	123	19	94	42	136	10	103	29
Naive Bayes	106	14	99	59	119	5	108	46
Decision Tree	161	84	29	4	162	79	34	3
K-NN	142	39	74	23	151	33	80	14
Random Forest	159	79	34	6	162	68	45	3

Table-II shows the confusion matrix that will be used to calculate performance accuracy, precision, recall and f measure. It compares the confusion matrix of sentiment analysis using TF-IDF, and sentiment analysis using combination of TF-IDF and Backward Elimination in five different machine learning methods.

Table-III: Results Comparison Without Feature Selection

Method	TF-IDF				
	Accuracy	Precision	Recall	F measure	Runtime
SVM	78.06%	69.12%	83.19%	75.50%	2 s
Naive Bayes	73.74%	62.66%	87.61%	73.06%	1 s
Decision Tree	68.35%	87.88%	25.66%	39.73%	1 s
K-NN	77.70%	76.29%	65.49%	70.48%	1 s
Random Forest	69.42%	85.00%	30.09%	44.44%	1 s

Table-IV: Results Comparison With Feature Selection

Method	TF-IDF & Backward Elimination				
	Accuracy	Precision	Recall	F measure	Runtime
SVM	85.97%	78.03%	91.15%	84.08%	41 s
Naive Bayes	81.65%	70.13%	95.58%	80.90%	4 s
Decision Tree	70.50%	91.89%	30.09%	45.33%	22 s
K-NN	83.09%	85.11%	70.80%	77.29%	58 s
Random Forest	74.46%	93.75%	39.82%	55.90%	123 s

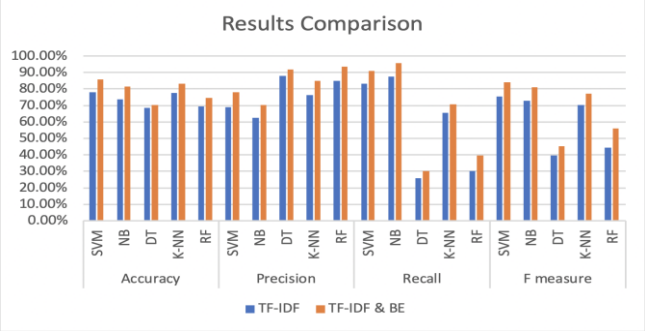

Fig 5. Results Comparison

Table-III and Table-IV shows the results comparison of performance accuracy, precision, recall, and f measure between sentiment analysis using TF-IDF and sentiment analysis using combination of TF-IDF and Backward Elimination in five different machine learning methods.

The highest accuracy for classifying sentiments in this research is 85.97%. It is achieved by using SVM and combination of TF-IDF and Backward Elimination. Although Backward Elimination feature selection increases the process runtime, it has shown better results in performance accuracy, precision, recall, and f measure for all classifiers used in this paper. Therefore Backward Elimination feature selection succeeded in achieving the expectation of this research.

The highest accuracy for classifying sentiments in this research is 85.97%. It is achieved by using SVM and combination of TF-IDF and Backward Elimination.

Although Backward Elimination feature selection increases the process runtime, it has shown better results in performance accuracy, precision, recall, and f measure for all classifiers used in this paper. Therefore Backward Elimination feature selection succeeded in achieving the expectation of this research. The results of this research shows that feature selection method can be an option to improve the performance in sentiment analysis.

V. CONCLUSION

The objective of this research is to optimize sentiment analysis performance by using feature selection strategy. Product reviews from Tokopedia, Shopee, and Bukalapak was used as the dataset, while TF-IDF feature extraction, Backward Elimination feature selection, and SVM, Naive Bayes, Decision Tree, K-NN, Random Forest classifiers was used in analysing the sentiment of the reviews. The best accuracy is achieved by using TF-IDF and Backward Elimination in SVM with a score of 85.97%, which increases by 7.91% after applying feature selection. From the results, Backward Elimination succeeded in improving all performance including accuracy, precision, recall, and f measure for all classifiers used in this research if compared to sentiment analysis that did not use any feature selection. The concern in using Backward Elimination feature selection is longer runtime when dataset gets bigger. Overall, it can be concluded that feature selection technique can be used to optimize performance of 2 class classification in sentiment analysis on e-commerce product reviews. For future works in this research, it is highly recommendable to use larger datasets and to do comparison with other feature selection methods.

ACKNOWLEDGMENT

The authors would like take this opportunity to express their deepest gratitude to all those who have helped in completing this study, especially to Bina Nusantara University for supporting this research project.

REFERENCES

1. N. Kristiadi, "E-Commerce, Manfaat, dan Keuntungannya," 15 August 2017. [Online]. Available: https://www.kompasiana.com/novikristiadi/5992634e93be2508e06c5402/e-commerce-manfaat-dan-keuntungannya. [Accessed 13 November 2019].
2. Aseanup, "Top 10 e-commerce sites in Indonesia 2019," 6 November 2019. [Online]. Available: https://aseanup.com/top-e-commerce-sites-indonesia/. [Accessed 13 November 2019].
3. R. Liang and J.-q. Wang, "A Linguistic Intuitionistic Cloud Decision Support Model with Sentiment Analysis for Product Selection in E-commerce," International Journal of Fuzzy System, 2019.
4. T. U. Haque, N. N. Saber and F. M. Shah, "Sentiment Analysis on Large Scale Amazon Product Reviews," IEEE International Conference on Innovative Research and Development, 2018.
5. R. Safrin, K. Sharmila, T. Subangi and E. Vimal, "Sentiment Analysis on Online Product Review," International Research Journal of Engineering and Technology(IRJET), vol. 4, no. 4, pp. 2381-2388, 2017.
6. A. N. Kamilah, "Analisa Sentimen Pelanggan Tokopedia Menggunakan Algoritma Naive Bayes Berdasarkan Review Pelanggan," Simki-Techsain, vol. 1, no. 6, pp. 1-13, 2017.
7. H. L. Adi, E. Sediyono and R. Kusumaningrum, "Support Vector Machine Classifier for Sentiment Analysis of Feedback Marketplace with a Comparison Features at Aspect Level".

8. B. Gunawan, H. S. Pratiwi and E. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," Jurnal Edukasi dan Penelitian Informatika(JEPIN), vol. 4, no. 2, pp. 113-118, 2018.
9. M. D. P. Artanti, A. Syukur, A. Prihandono and D. R. I. M. Setiadi, "Analisa Sentimen untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes," in Konferensi Nasional Sistem Informasi 2018, Pangkalpinang, 2018.
10. K. Sudheer and B. Valarmathi, "Real Time Sentiment Analysis of E-Commerce Websites Using Machine Learning Algorithms," International Journal of Mechanical Engineering and Technology(IJMET), vol. 9, no. 2, pp. 180-193, 2018.
11. Y. Hedge and S. Padma, "Sentiment Analysis Using Random Forest Ensemble for Mobile Product Reviews in Kannada," IEEE 7th International Advance Computing Conference, pp. 777-782, 2017.
12. M. R. Herga, "Implementasi Text Mining Sistem Klasifikasi dan Pencarian Konten Buku Perpustakaan Menggunakan Algoritma Naïve Bayes Classifier".
13. D. Virmani and S. Taneja, "A Text Preprocessing Approach for Efficacious Information Retrieval," Smart Innovations in Communication and Computational Sciences, Advances in Intelligent Systems and Computing 669, pp. 13-22, 2019.
14. M. A. Sghaier and M. Zrigui, "Sentiment Analysis for Arabic E-commerce Websites," in 2016 International Conference on Engineering & MIS (ICEMIS), Agadir, Morocco, 2016.
15. O. Somantri, "Text Mining untuk Klasifikasi Kateogori Cerita Pendek Menggunakan Naïve Bayes (NB)," Jurnal Telematika, vol. 12, no. 1, 2017.
16. V. Kalra and R. Aggarwal, "Importance of Text Data Preprocessing & Implementation in RapidMiner," in First International Conference on Information Technology and Knowledge Management (ICTKM), New Delhi, 2018.
17. Y. T. Pratama, F. A. Bachtiar and N. Y. Setiawan, "Analisis Sentimen Opini Pelanggan Terhadap Aspek Pariwisata Pantai Malang Selatan Menggunakan TF-IDF dan Support Vector Machine," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 2, no. 12, pp. 6244-6252, 2018.
18. S. Fatima and B. Srinivasu, "Text Document Categorization Using Support Vector Machine," International Research Journal of Engineering and Technology(IRJET), vol. 4, no. 2, pp. 141-147, 2017.
19. "Feature Extraction," DeepAI, [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/feature-extraction. [Accessed 8 February 2020].
20. B. Kurniawan, S. Effendi and O. S. Sitompul, "Klasifikasi Konten Berita dengan Metode Text Mining," Jurnal Dunia Teknologi Informasi, vol. 1, no. 1, pp. 14-19, 2012
21. P. Kumbhar and M. Mali, "A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification," International Journal of Science and Research(IJSR), vol. 5, no. 5, pp. 1267-1275, 2016.
22. "Backward Elimination(RapidMiner Studio Core)," [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/modeling/optimization/feature_selection/optimize_selection_backward.html. [Accessed 8 February 2020].
23. M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez and K. Kochut, "A Brief Survey of Text Mining: Classification, Clustering, and Extraction Techniques," arXiv, 2017.
24. A. A. Lutfi, A. E. Permanasari and S. Fauziati, "Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine," Journal of Information Systems Engineering and Business Intelligence, vol. 4, no. 1, pp. 57-64, 2018.
25. M. N. Saadah, R. W. Atmagi, D. S. Rahayu and A. Z. Arifin, "Information Retrieval of Text Document with Weighting TF-IDF and LCS," Journal of Computer Science and Information, vol. 6, no. 1, p. 34, 2013.
26. K. M. Almunirawi and A. Y. Maghari, "A Comparative Study on Serial Decision Tree Classification Algorithms in Text Mining," International Journal of Intelligent Computing Research(IJICR), vol. 7, no. 4, pp. 754-760, 2016.
27. A. Sukma, B. Zaman and E. Purwanti, "Information Retrieval Document Classified with K-Nearest Neighbor," Record and Library Journal, vol. 1, no. 2, pp. 129-138, 2015.
28. Bahrawi, "Sentiment Analysis Using Random Forest Algorithm - Online Social Media Based," Journal of Information Technology and Its Utilization, vol. 2, no. 2, pp. 29-33, 2019.
29. S. Narkheda, "Understanding Confusion Matrix," 9 May 2018. [Online]. Available: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62. [Accessed 5 February 2020].
30. P. Dellia and A. Tjahyanto, "Tax Complaints Classification on Twitter Using Text Mining," Journal of Science, vol. 2, no. 1, pp. 11-15, 2017.

AUTHORS PROFILE

Tommy Willianto is a graduate student in Computer Science Department, Bina Nusantara University, BINUS Graduate Program-Master in Computer Science, Jakarta, 11480, Indonesia. From September 2017 – December 2018, he worked in a software house as a software engineer. In March 2019 – August 2019, he had his internship in a technology based startup as a backend engineer. He starts having interest in research since early 2019. Afterwards, he decided to change his career path from an engineer to a researcher. Currently he chooses to focus in the field of data mining on his research. Other research topics that he is interest in are big data analytics, business intelligence, internet of things and optimization.

Supryadi is a graduate student in Computer Science Department, Bina Nusantara University, BINUS Graduate Program-Master in Computer Science, Jakarta, 11480, Indonesia. In 2019, he had his 6 month internship in software and IT company with a position as a system analyst. His research interest in text mining, data warehouse and machine learning.

Antoni Wibowo has received his first degree of Applied Mathematics in 1995 and a master degree of Computer Science in 2000. In 2003, He was awarded a Japanese Government Scholarship (Monbukagakusho) to attend Master and PhD programs at Systems and Information Engineering in University of Tsukuba-Japan. He completed the second master degree in 2006 and PhD degree in 2009, respectively. His PhD research focused on machine learning, operations research, multivariate statistical analysis and mathematical programming, especially in developing nonlinear robust regressions using statistical learning theory. He has worked from 1997 to 2010 as a researcher in the Agency for the Assessment and Application of Technology – Indonesia. From April 2010 – September 2014, he worked as a senior lecturer in the Department of Computer Science - Faculty of Computing, and a researcher in the Operation Business Intelligence (OBI) Research Group, Universiti Teknologi Malaysia (UTM) – Malaysia. From October 2014 – October 2016, he was an Associate Professor at Department of Decision Sciences, School of Quantitative Sciences in Universiti Utara Malaysia (UUM). Dr. Eng. Wibowo is currently working at Binus Graduate Program (Master in Computer Science) in Bina Nusantara University-Indonesia as a Specialist Lecturer and continues his research activities in machine learning, optimization, operations research, multivariate data analysis, data mining, computational intelligence and artificial intelligence.