Bulletin of Electrical Engineering and Informatics
Vol. 9, No. 5, October 2020, pp. 2012~2019
ISSN: 2302-9285, DOI: 10.11591/eei.v9i5.2444

Indonesian language email spam detection using N-gram and Naïve Bayes algorithm

Yustinus Vernanda, Seng Hansun, Marcel Bonar Kristanda
Department of Informatics, Universitas Multimedia Nusantara, Indonesia

Article Info

Article history:

Received Feb 20, 2020
Revised Apr 3, 2020
Accepted May 14, 2020

Keywords:

Email
N-gram
Naïve Bayes
Spam filter
Web service

ABSTRACT

Indonesia is ranked the top 8th out of the total country population in the world for the global spammers. Web-based spam filter service with the REST API type can be used to detect email spam in the Indonesian language on the email server or various types of email server applications. With REST API, then there will be data exchange between the applications with JSON data type using existing HTTP commands. One type of spam filter commonly used is Bayesian Filtering, where the Naïve Bayes algorithm is used as a classification algorithm. Meanwhile, the N-gram method is used to increase the accuracy of the implementation of the Naïve Bayes algorithm in this study. N-gram and Naïve Bayes algorithms to detect spam email in the Indonesian language have successfully been implemented with accuracy around 0.615 until 0.94, precision at 0.566 until 0.924, recall at 0.96 until 1.00, and F-measure at 0.721 until 0.942. The best solution is found by using the 5-gram method with the highest score of accuracy at 0.94, precision at 0.924, recall at 0.96, and F-measure value at 0.942.

Corresponding Author:

Seng Hansun,
Department of Informatics,
Universitas Multimedia Nusantara,
Jl. Scientia Boulevard, Gading Serpong, Tangerang, Banten-15811 Indonesia.
Email: seng.hansun@lecturer.umn.ac.id

1. INTRODUCTION

Spam is an unsolicited email that is sent to the crowd [1]. According to Suryanto [2], the number of spam emails in the world increases exponentially every year. Based on recent spam statistics data from AV-test, Indonesia is ranked 8th of the total country population in the world for the global spammers [3]. The regulations on spam spreading in Indonesia have not been explicitly regulated in the Information and Transaction Act Electronic (Law No. 11 Year 2008/UU ITE). However, spam delivery can be categorized in deeds is forbidden in Chapter VII, articles 27-34, more precisely chapter 33 [4].

Various researches have been done related to spam detection and filter as we can see in the works of Nagwani and Sharaff [5], Sah and Parmar [6], Bhuiyan et al. [7], Ezpeleta et al. [8], and Jawale et al. [9]. However, the most used method to prevent spam is a text mining method with Bayesian filtering. Even though many advanced text mining techniques have been developed [10-15] and comparison between different methods has been done [16-18], the Naïve Bayes algorithm is considered simple and has a fast computation [19, 20]. Moreover, the N-gram method is used to add the accuracy of the Naïve Bayes algorithm inside the spam classifier, as we can see in [21-23].

Web service is defined as an interface that describes a set of operations that can be accessed through the network [24]. Web service usage aims to be used on mail servers and mail clients on various platform types [25]. The most used protocol to access API is REST [26]. The main advantage of

using REST is the bandwidth used is less than SOAP because SOAP requires XML wrapper for every request and response [25]. Text mining can be used to handle problems of classification, clustering, and information extraction and retrieval [27]. Text mining and data mining are differed from the source used. The data source used in data mining is structured data, while the data source used on text mining is unstructured data in text form [4]. The initial stage in text mining is text pre-processing, i.e., the process of changing the form of data not yet structured into structured data according to needs, which are done for more mining processes to continue. The steps in the pre-processing text, in general, are case-folding, tokenizing, filtering, and stemming [27]. Case-folding is a process to change all the characters in the document into lowercase [28]. Tokenizing is the cutting stage of input text into words, terms, symbols, punctuation, or another element that has a meaning called a token [29]. Filtering is the stage of picking up essential words of results token [30]. Stemming is the process of mapping variance morphological words in the base or general word [31].

In this research, we are trying to detect spam email in the Indonesian language using a text mining method, namely the Naïve Bayes classifier, which was enhanced with the N-gram method. Different from previous research, in this study, we propose and implement both Naïve Bayes and N-gram methods as a web service using REST API design. Further information on those methods and REST API design is given in the following section. Section 3 delivers the implementation results and analysis of spam detection results by calculating the accuracy, precision, recall, and F-measure. In the end, some concluding remarks will be given in section 4.

2. RESEARCH METHOD

2.1. N-gram

A set of n-character which is taken from a string is called N-gram [32]. The N-gram method was used for taking pieces of capital letters in a continuous word from the source until the end of the string. If n=1 then it's a unigram, if n=2 it's a bigram, and if n=3 it's a trigram. For example, the word "bagus" can be formed into several N-gram as:

- Unigram : b, a, g, u, s
- Bigram : _b, ba, ag, gu, us, s_
- Trigram : _ba, bag, agu, gus, us_, s__

Blank "_" character is used to represent space on the beginning and on the end of the word. The advantage of using N-gram is based on the characteristics of the N-gram as a part of a string, so the error on a partial string only resulting in a difference in some N-gram [33]. Another representation and usage of N-gram advantage can also be seen in the publication of Tayyeh and Al-Jumaili [34].

2.2. Naïve Bayes

Naïve Bayes algorithm is advanced by English scientists, Thomas Bayes. This algorithm utilizes the probability and statistics methods to predict probability in the future based on past experience [35]. It requires a small amount of data training [36], and the basis of the Naive Bayes theorem used in programming is the following Bayes formula [37].

In (1) shows the probability of occurrence of A when B determined from the probability B when A, probability A, and probability B. Naive Bayes classifier or may be referred to as multinomial Naive Bayes is a simplified model of the Bayes algorithm that fits in the classification of text or documents by (2) like the following [35].

$$P\ (A|B) = (P(B|A)\ *\ P(A))/P(B) \quad (1)$$

$$V_{MAP} = arg\ \max_{V_j \in V} \prod_{i=1}^{n} P\left(x_i \middle| V_j\right) P(V_j) \quad (2)$$

where: V_{MAP} = Category or class that has the highest posterior
V_j = Category or class j=1, 2, 3, ..., n
x_i = Words, i=1, 2, 3, ..., n
$P\left(x_i \middle| V_j\right)$ = Probability x_i in category V_j
$P(V_j)$ = Probability of V_j
$argmax$ = Domain that has the greatest value
$V_j \in V = V_j$ = The element or set of V

For $P(V_j)$ and $P\left(x_i \middle| V_j\right)$ calculated by (3) and (4) as follows [35].

$$P(V_j) = \frac{|\,docs_j\,|}{|\,sample\,|} \quad (3)$$

$$P(x_i|V_j) = \frac{n_k+1}{n+|words|} \quad (4)$$

where: $|\,docs_j\,|$ = Total number of documents in j
$|\,sample\,|$ = Total number of documents
n_k = Number of occurrence for each word
n = Number of word occurrence for each category
$|words|$ = Total number of words from all categories

2.3. REST API

The term REST which stands for representational state transfer was first used by Roy Thomas Fielding, one of the pioneers of the Apache webserver project, in his doctoral dissertation at the University of California in 2000 [38]. REST API architecture components are client applications, networks, and web services. The client application sends an HTTP request containing methods like GET, PUT, POST and others to web services over the network. Design of application programming interface (API) includes the design method and design of JSON. The design method is used to design the URI request pattern and what type of method is used to send HTTP requests. JSON design is used to design JSON data that is sent to the client.

Figure 1 shows the API flow spam filter created in this system. There are seven functions that can be sent by the client application to web service through an API gateway. The seven functions are user key, remove user key, user list, add dataset, list dataset, remove dataset, and check spam. The usefulness of the seven functions is explained in the design method. Client application sends request via HTTP request using specified method. When sending an HTTP request, the client application must use authentication. Authentication that is sent by the client application is the header with parameter "X-API-KEY" filled by key owned by each client application. After the client application sends the request, the web service responds in the form of JSON to the client application. The results of these responses are then managed by the client application in accordance with their needs.

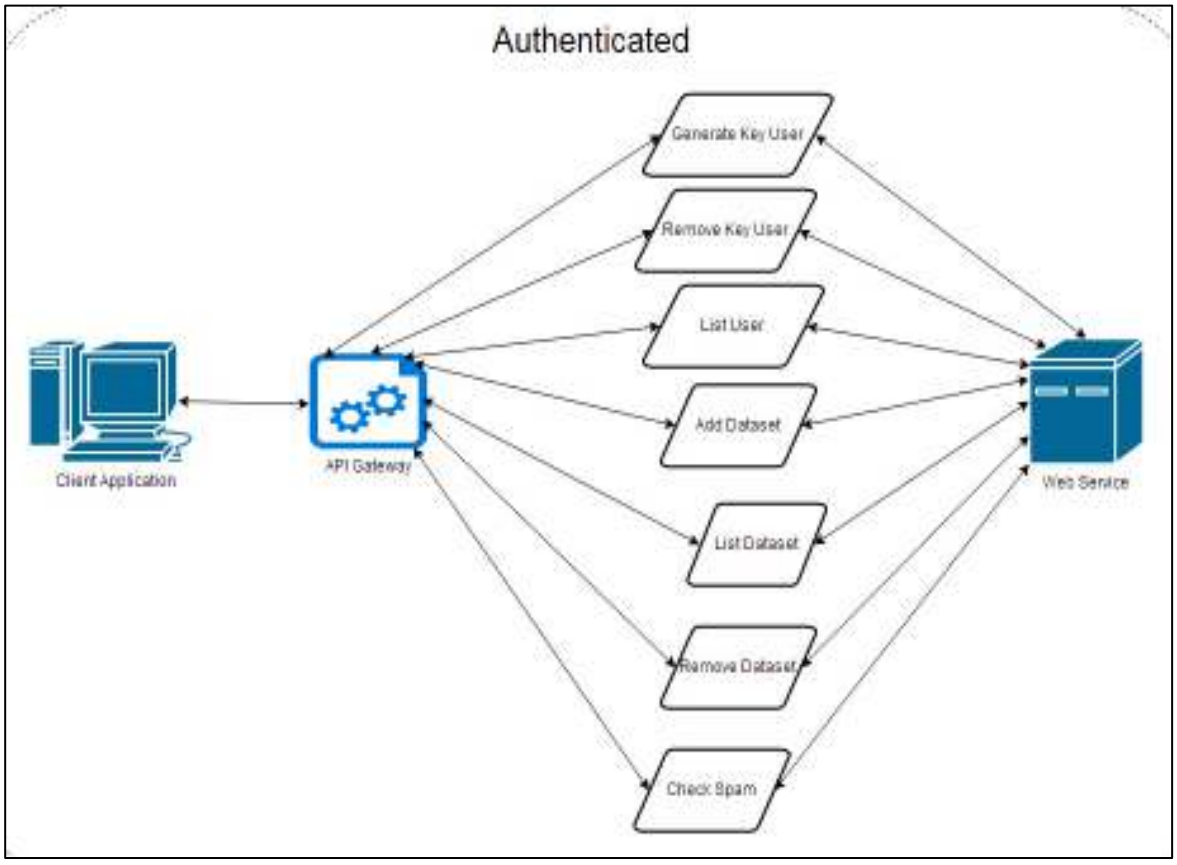

Figure 1. API flow of spam filter

3. RESULTS AND DISCUSSION

Based on the design, there is seven main design of the application interface, namely content page, visitor menu, user menu, sign in menu, register menu, console menu, admin page. Figure 2 shows the implementation of the content page design. In accordance with the design, there are headers, sidebar, footer, and content columns consisting of four components of explanation column, input form field, try button, and result field. Implementation of content page design is used in case folding, tokenizing, filtering, stemming, and N-gram. Figure 2 is the content page of the case folding menu. In the case of case-folding, there is a form to enter a sentence. The sentence is used to simulate the result of the case-folding process. When a visitor or user is pressing the try button, then the system to process the case folding and display the results.

Figure 2. The content page

Figure 3 is an implementation of the console menu design. In the console menu there are forms, sent buttons, and results columns. The form in the console menu consists of an input form to include HTTP requests, API keys, and body parameters. When the user pressed the send button, the system runs the HTTP request entered by the user. Then the result field displays the JSON result of the sent request.

Figure 3. Implementation of menu console design

Figure 4 shows the implementation of the admin page design. In accordance with the design, on the admin page consists of a header that consists of the application name and sign out menu. Then the content column on the admin page displays user information. For each single user information, there are button suspend, active, and upgrade. If the suspend button is pressed then the user cannot use an HTTP request. If the active button is pressed, then the user can use an HTTP request. If the upgrade button is pressed then the user can use an HTTP request without any limit.

Spam filter trial is done by several methods of measurement, which are accuracy, precision, recall, and f-measure. The values obtained from the calculation ranges from 0 to 1, where higher value means better result, and vice versa. Spam filter testing uses 100 spam category documents and 100 ham category documents with training data of 200 spam-category documents and 200 ham category documents. URI dataset contains all data used as training and test data in this study [39]. The total amount of training and test data in this study is 600 documents which were obtained from 30 people. Everyone gives ten spam emails and ten ham emails. The first 20 data were used as training data while the last ten data were used as test data.

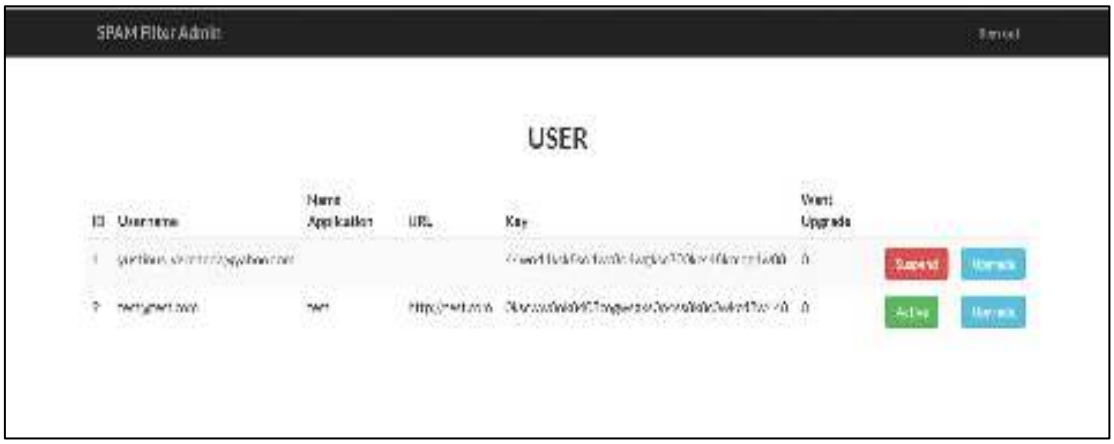

Figure 4. Implementation of admin page design

Table 1 shows the test results for spam filters on each N-gram method. Based on the test results, the lowest accuracy, precision, and f-measure value is spam filter using the 1-gram method with an accuracy value of 0.615, a precision value of 0.566, and an f-measure value of 0.721. While the lowest recall value is spam filter using the 5-gram method until the 10-gram method with accuracy value equal to 0.96. Then for the highest accuracy, recall, and f-measure value is spam filter using the 5-gram method with an accuracy value 0.94, a precision value equal to 0.924, and an f-measure value 0.942. Meanwhile, the highest recall value is a spam filter that uses the 2-gram method with a value of 1.

Table 1. Results of spam filter

N-gram	Accuracy	Recall	Precision	F-measure
0	0.935	0.97	0.907	0.938
1	0.615	0.99	0.566	0.721
2	0.64	1	0.582	0.736
3	0.695	0.99	0.623	0.765
4	0.89	0.97	0.837	0.899
5	**0.94**	**0.96**	**0.924**	**0.942**
6	0.935	0.96	0.915	0.937
7	0.935	0.96	0.915	0.937
8	0.935	0.96	0.915	0.937
9	0.935	0.96	0.915	0.937
10	0.935	0.96	0.915	0.937

Figure 5 shows a graph of test results on a spam filter. In the graph of test results on spam filters, it can be said that the 6-gram method onwards does not lead to significant changes in the implementation of the Naive Bayes algorithm. The argument is obtained based on the analysis of accuracy, precision, recall, and f-measure values that do not change or stable on the 6-gram method until 10-gram with a precision

value equal to 0.915, a recall value equal to 0.96, an accuracy value equal to 0.935, and an f-measure value equal to 0.937.

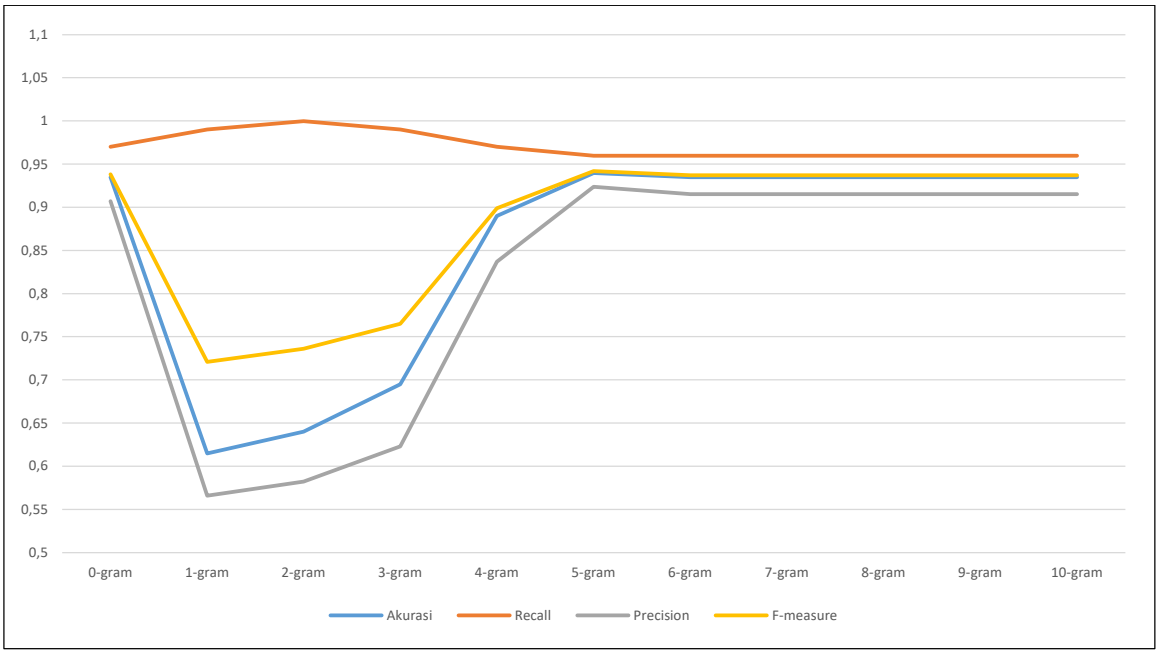

Figure 5. Spam filter test graph

4. CONCLUSION

In this study, the N-gram method and Naïve Bayes algorithm had been successfully implemented to detect Indonesian language spam using REST API architecture. From the experimental results, it can be concluded that the accuracy values ranged from 0.615 to 0.94, the precision values ranged from 0.566 to 0.924, the recall values ranged from 0.96 to 1, and the f-measure values ranged from 0.721 to 0.942. The 6-gram method and later did not have any significant change. Meanwhile, the best N-gram method that gives the highest accuracy, precision, and f-measure values in detecting Indonesian language spam is the 5-gram method when combined with the Naïve Bayes algorithm.

REFERENCES

[1] S. Salomon and S. Hansun, "Undergraduate student social media web spam filter using regular expression," in Bahasa "Spam filter situs jejaring sosial mahasiswa menggunakan regular expression," *ULTIMA InfoSys: Jurnal Ilmu Sistem Informasi*, vol. 8, no. 2, pp. 69-73, 2017.
[2] Suryanto, "Artificial intelligence searching, reasoning, planning and learning," Informatika, Bandung, 2014.
[3] AV-TEST-The Independent IT-Security Institute, "Spam statistics & trends report," 2020. [Online]. Available at: https://www.av-test.org/en/statistics/spam/.
[4] C. A. Nugroho, Samsudi, and D. E. Nurhayati, "Penal policy about distributing spam via short message service," Thesis, Universitas Jember (UNEJ), Indonesia, 2013.25], [28],
[5] N. K. Nagwani and A. Sharaff, "SMS spam filtering and thread identification using bi-level text classification and clustering techniques," *Journal of Information Science*, vol. 43, no. 1, pp. 75-87, 2017.
[6] U. K. Sah and N. Parmar, "An approach for malicious spam detection in email with comparison of different classifiers," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 8, pp. 2238-2242, 2017.
[7] H. Bhuiyan, A. Ashiquzzaman, T. I. Juthi, S. Biswas, and J. Ara, "A survey of existing e-mail spam filtering methods considering machine learning techniques," *Global Journal of Computer Science and Technology*, vol. 18, no. 2, pp. 20-29, 2018.
[8] E. Ezpeleta, I. Garitano, U. Zurutuza, and J. M. G. Hidalgo, "Short messages spam filtering combining personality recognition and sentiment analysis," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 25, no. 2, pp. 175-189, 2017.
[9] D. S. Jawale, A. G. Mahajan, K. R. Shinkar, and V. V. Katdare, "Hybrid spam detection using machine learning," *Int. Journal of Advance Research, Ideas and Innovations in Technology*, vol. 4, no. 2, pp. 28-28-2832, 2018.

[10] L. Andersson, A. Hanbury, and A. Rauber, "The portability of three types of text mining techniques into the patent text genre," *Current Challenges in Patent Information Retrieval*, (M. Lupu, K. Mayer, N. Kando, A. Trippe eds.), *The Information Retrieval Series*, vol. 37, pp. 241-280, 2017.
[11] D. Yu, Z. Xu, W. Pedrycz, and W. Wang, "Information sciences 1968-2016: A retrospective analysis with text mining and bibliometric," *Information Sciences*, vol. 418-419, pp. 619-634, 2017.
[12] V. Kayser and K. Blind, "Extending the knowledge base of foresight: The contribution of text mining," *Technological Forecasting and Social Change*, vol. 116, pp. 208-215, 2017.
[13] R. A. Saravanan and M. R. Babu, "Enhanced text mining approach based on ontology for clustering research project selection," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-11, 2017.
[14] S. Redhu, S. Srivastava, B. Bansal, and G. Gupta, "Sentiment analysis using text mining: A review," *International Journal on Data Science and Technology*, vol. 4, no. 2, pp. 49-53, 2018.
[15] K. Marimuthu, A. Shankar, R. Ranganathan, and R. Niranchana, "Product opinion analysis using text mining and analysis," *International Journal of Smart Grid and Green Communications*, vol. 1, no. 3, pp. 227-234, 2018.
[16] Md. Shoeb and J. Ahmed, "Sentiment analysis and classification of tweets using data mining," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 12, pp. 1471-1474, 2017.
[17] W. B. Zulfikar, M. Irfan, C. N. Alam, and M. Indra, "The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter," *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, pp. 1-5, 2017.
[18] F. R. Lucini, F. S. Fogliatto, G. J. C. da Silveira, J. L. Neyeloff, M. J. Anzanello, R. S. Kuchenbecker, and B. D. Schaan, "Text mining approach to predict hospital admissions using early medical records from the emergency department," *International Journal of Medical Informatics*, vol. 100, pp. 1-8, 2017.
[19] S. Das and A. K. Kolya, "Sense GST: Text mining & sentiment analysis of GST tweets by Naive Bayes algorithm," *2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pp. 239-244, 2017.
[20] V. Ferdina, M. B. Kristanda, and S. Hansun, "Automated complaints classification using modified Nazief-Adriani stemming algorithm and Naive Bayes classifier," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 5, pp. 1604-1614, 2019.
[21] S. Ahmad and R. Varma, "Information extraction from text messages using data mining techniques," *Malaya Journal of Matematik*, vol. S, no. 1, pp. 26-29, 2018.
[22] M. Schonlau, N. Guenther, and I. Sucholutsky, "Text mining with n-gram variables," *The Stata Journal*, vol. 17, no. 4, pp. 866-881, 2017.
[23] N. V. Mathew and V. R. Bai, "Analyzing the effectiveness of N-gram technique based feature set in a Naive Bayesian spam filter," *Proc. of Int. Conference on Emerging Technological Trends (ICETT)*, pp.1-5, 2016.
[24] H. Kreger, "Web services conceptual architecture (WSCA 1.0)," IBM Software Group, USA, 2001.
[25] E. Kurniawan, "REST web service implementation for mobile based sales order and sales tracking," in Bahasa "REST web service untuk sales order dan sales tracking berbasis mobile," *Jurnal EKSIS*, vol.7, no.1, pp. 1-12, 2014.
[26] A. DuVander, "The next wave? Enterprises moving SOAP to REST," Cicero API, 2012. [Online]. Available at: https://www.programmableweb.com/news/next-wave-enterprises-moving-soap-to-rest/2012/03/22.
[27] M. W. Berry and J. Kogan, "Text mining: Applications and theory," John Wiley & Sons, United States, 2010.
[28] C. Triawati, "Statistical concept based weighting method for clustering and categorization of Indonesian Language document," in Bahasa "Metode pembobotan statistical concept based untuk klastering dan kategorisasi dokumen berbahasa Indonesia," Thesis, TELKOM University, 2009.
[29] S. Vijayarani and R. Janani, "String matching algorithms for retrieving information from desktop-Comparative analysis," *2016 International Conference on Inventive Computation Technologies (ICICT)*, pp.1-6, 2016.
[30] R. J. Mooney, "CS 391L: Machine learning text categorization," Presentation, University of Texas at Austin, 2006. [Online]. Available at: https://www.cs.utexas.edu/~mooney/cs391L/slides/text.pdf.
[31] R. Adhitia and A. Purwarianti, "Indonesian Language essay scoring using SVM-LSA method with generic feature," in Bahasa "Penilaian esai jawaban bahasa Indonesia menggunakan metode SVM-LSA dengan fitur generic," *Journal of Information System*, vol.5, no.1, pp. 33-41, 2009.
[32] Y. Permadi, "Text categorization using N-gram for documents in Indonesian Language," in Bahasa "Kategori teks menggunakan N-gram untuk dokumen berbahasa Indonesia," Thesis, Institut Pertanian Bogor, 2008.
[33] W. B. Cavnarand and J. M. Trenkle, "N-gram-based text categorization," Environmental Research Institute of Michigan. [Online]. Available at: https://www.let.rug.nl/vannoord/TextCat/textcat.pdf.
[34] H. K. Tayyeh and A. S. A. Al-Jumaili, "Classifying confidential data using SVM for efficient cloud query processing," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 17, no. 6, pp. 3155-3160, 2019.
[35] P. Graham, "A plan for spam," 2002. [Online]. Available at: http://www.paulgraham.com/spam.html.
[36] W. N. L. W. H. Ibeni, M. Z. M. Salikon, A. Mustapha, S. A. Daud, and M. N. M. Salleh, "Comparative analysis on Bayesian classification for breast cancer problem," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, pp. 1303-1311, 2019.
[37] A. Fadil, I. Riadi, and S. Aji, "Review of detection DDOS attack detection using Naïve Bayes classifier for network forensics," *Bulletin of Electrical Engineering and Informatics*, vol. 6, no. 2, pp. 140-148, 2017.
[38] R. T. Fielding, "Architectural styles and the design of network-based software architectures," Doctoral Dissertation, University of California, California, 2000.
[39] "URI dataset," [Online]. Available at: https://tinyurl.com/datasetTA.

BIOGRAPHIES OF AUTHORS

Yustinus Vernanda was born on June 1st, 1995 in West Jakarta, Jakarta, Indonesia. He received his certificate in Certified Ethical Hacker on 2016. He completed his Computer Science bachelor degree from Universitas Multimedia Nusantara on 2017. In the same year, he worked in PT. Pratapa Nirmala Company, one of pharmacy companies in Indonesia. His research interests lie in the application of big data to support network security.

Seng Hansun lives in Tangerang, Indonesia. He received his Bachelor degree in Mathematics and his Master degree in Computer Science from Universitas Gadjah Mada, Yogyakarta. In 2011, he began his academic experience as a Lecturer in Computer Science Department of Universitas Multimedia Nusantara. He has published one text book in Android Programming and more than 100 articles both nationally and internationally. Computational science, soft computing methods, internet, mobile technology, and medical informatics are some of his research interests lately.

Marcel Bonar Kristanda has received his Bachelor degree in Computer Science from Universitas Multimedia Nusantara, Indonesia, in 2011 and his Master degree from Chinese Culture University, Taiwan, in 2015. He began his career as an Assistant Lecturer in 2011, and now has become a Lecturer in Computer Science Department, Universitas Multimedia Nusantara. He also was entrusted to lead the Learning Center Department in 2016 at UMN. His research mainly based on his interests in mobile technology and application development, web development, and software engineering.