RESEARCH ARTICLE OPEN ACCESS

Manuscript received December 31, 2022; revised January 21, 2023; accepted February 02, 2023; date of publication February 25, 2023
Digital Object Identifier (**DOI**): https://doi.org/10.35882/ijeeemi.v5i1.262

How to cite: Annisa Puspa Kirana, Gunawan Budi Prasetyo, and Ela Widya Lestari, "Detection of Indonesian Hoax Content about COVID-19 Vaccine using Naive Bayes Multinomial Method", Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 5, no. 1, pp. 13–19, February. 2023.

Detection of Indonesian Hoax Content about COVID-19 Vaccine using Naive Bayes Multinomial Method

Annisa Puspa Kirana, Gunawan Budi Prasetyo, and Ela Widya Lestari
Information Technology Department, State Polytechnic of Malang, Indonesia

Corresponding author: Annisa Puspa Kirana (email: puspakirana@polinema.ac.id).

ABSTRACT One media currently famously used in all worlds is Twitter. The ease of dissemination and the exchange of information is accelerating. Every day, millions of tweets exist using various information, such as politics, technology, sports, academics, and others. The information that is widely found is about COVID-19-19 nowadays. The information on Twitter is not entirely accurate or according to facts and needs to be proven true. Therefore, this study aims to try to detect the information contained in Indonesia using methods of Naive Bayes Multinomial by using the Information Gain feature selection. This research contributes to utilizing data spread on Twitter and social media in detecting hoaxes spread in the community, primarily related to COVID-19 infections. The classification process is carried out by crawling tweets, preprocessing, then using feature selection, namely Information Gain, and classification using the Multinomial Naive Bayes method. Meanwhile, the validation needs in this study use k-fold cross-validation where the existing dataset is divided into training and testing data that will be tested with a confusion matrix. Researchers have carried out the confusion matrix testing process using 720 datasets divided as train data & the test data received an average accuracy value of 81.39%, precision of 80.36%, and recall of 79.73%. The highest accuracy is using k-fold two. The accuracy value reaches 88.8%, the precision value is 79.1%, and the recall value is 86.3%. The lowest accuracy was obtained on the 8th k-fold with an accuracy value of 73.6%, a result precision of 75.4%, and a recall of 86.9%.

INDEX TERMS text mining, information gain, naive Bayes, multinomial, Twitter, covid-19

I. INTRODUCTION

Social media is a means to exchange text, images, video, or audio messages. Currently, Indonesia's active social media users are getting higher every year. Indonesia already has one social media user, one of which is Twitter. Each user is free to upload tweets containing both positive and negative content. COVID-19 is the information that is widely found today. COVID-19 information circulating on social media is not entirely accurate or according to facts; information that is not true or false can be called a hoax. To stop the spread of COVID-19, the Indonesian government will also vaccinate Indonesian residents. News lies about the composition of COVID-19 vaccines, hoax news about the impact of COVID-19 vaccines, and hoax information about COVID-19 vaccines.

Currently, there is much research on detecting hoax news on Twitter. Several methods are often used in hoax news detection research on Twitter, such as the Support Vector Machine method [1]. Another study is Naïve Bayes was researched by Rahutomo et al. [2], Backpropagation was researched by Lhaksmana et al. [3], Naïve Forecast Method [12] and Prasetyo, Rino [4] studied Modified K-Nearest Neighbor. These methods have been successfully implemented and gained good accuracy. This good accuracy result depends on many things; apart from selecting the feature extraction method, it also depends on the type of feature used and the choice of the classification method.

However, the research conducted using that method also has some disadvantages. The Support Vector Machine method is limited to using a separator function that separates data into

two classes. When the class is divided into more than two, modification is needed when the training data is extensive, affecting the training time and memory size that will be necessary [5]. When we come across words in the test data for a specific class that isn't in the training data, we may end up with zero class probabilities. Meanwhile, this backpropagation method tends to slow to achieve convergence on receiving optimal accuracy and requires extensive data and optimization training that is used less efficiently. Therefore, it is necessary to improve optimally using other methods so that the results' accuracy is better, faster, and can be compared using only standard Backpropagation solving procedures [6][2].

With the rise of fake news, especially about the COVID-19 Vaccine, the author conducted a study to classify information on Twitter into hoaxes and non-hoaxes. The method used to calculate the frequency of each term of event data on a document and probability includes the Naive Bayes Multinomial Method using the Information Gain feature selection. The advantages of the Naive Bayes Multinomial include that there is already a high level of accuracy and easy to implement when computing is low. The minimum error rate and selection of Information Gain Features is a technique for reducing the number of appropriate or relevant features, then decreasing the dimensions of the elements in the data to be used [7]. Information Gain is an algorithm that functions as a limit determinant that will be used for available attributes; it can be only in 1 point or more than one attribute used, which symbolizes a reflection on the quality of a feature to be used [7].

This study aims to discover how to detect hoax news using Naive Bayes Multinomial and select the Information Gain feature. This study also discusses the accuracy results of seeing hoax news using Naive Bayes Multinomial using the Information Gain feature. This study used 720 Indonesian data taken by Twitter crawling, scraping from turnbackhoax.id websites. The data was taken from trending elements of hoaxes with the keyword "#vaksincovid19". Accuracy calculations use the performance of the confusion matrix to calculate accuracy, precision, and recall. Training data and test data to get the average accuracy results and what will be taken is the highest accuracy value. The contribution of this study is to create a system that can classify hoax news on Twitter, especially about the COVID-19 Vaccine. In addition, it utilizes data spread on Twitter, turnbackhoax.id websites, and communication and information technology in detecting hoaxes spread in the community.

II. PROPOSED METHOD

The Indonesian hoax content detection system for the covid-19 Vaccine is website-based. This system detects hoax news about the covid-19 Vaccine on Twitter using the information gain feature selection and the multinomial naïve Bayes method. This system is made for the surrounding community and students to be able to sort out the information circulating on social media as hoax or non-hoax news. The data used in this study is data taken crawlingly on Twitter using the hashtag #vaksincovid19 in Indonesian with a total of all data h 600 datasets. We also retrieved data from the turnbackhoax.id with a scraping process. We retrieved manually on the Ministry of Communication and Information Technology website a total of all data 720 data.

In this system, there are two users, namely admin and user, where the admin has full access rights to manage data and all activities in the detection system. In contrast, users only have access rights to view datasets, classification tests, and report hoaxes. This system is built based on a website using the Python programming language and uses flask as a medium to connect to the website using a code igniter framework. The data obtained will go through a manual labeling process, preprocessing word processing so the computer can read it correctly. The data is processed using the information gain feature option. Information gain is used to weight each word using and processed by the naïve classification method Bayes

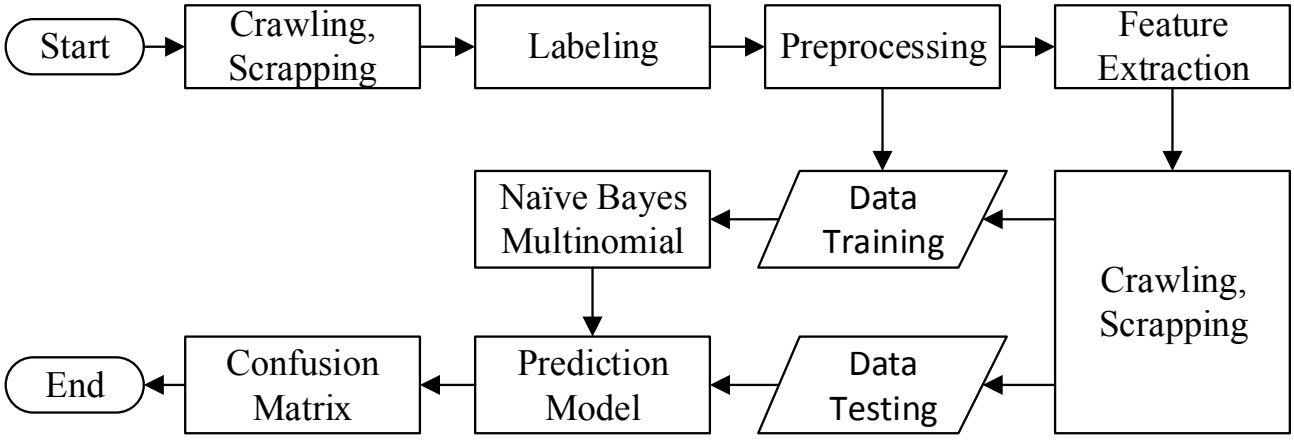


FIGURE 1. Design System

multinomial. This system uses a confusion matrix and k-fold cross-validation to measure the accuracy. The system can receive reports from the public and classify them with the models.

In this study, hoax news detection will be carried out on Twitter on topics related to #vaksincovid19. The system design used first is data crawling, labeling, preprocessing, feature extraction, and data features, then the data is divided into two, namely data training and data testing. Furthermore, the data train is processed using the multinomial naive Bayes method and then in the prediction model and confusion matrix. Meanwhile, the test data is directly processed into the prediction model and confusion matrix. FIGURE 1 is a proposed system method of this study.

A. CRAWLING

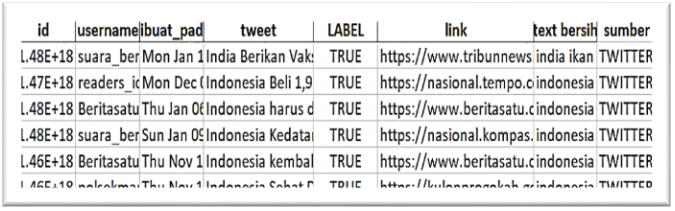
One of the data collection methods used for reference data by the system can be called data crawling [8]. In this study, the collection of APIs on Twitter. Every time you crawl your data, Twitter data can be set to a data limit. The data taken is a post based on keywords using hashtags that are estimated to contain hoax news. Furthermore, the results of crawling the crawling data will be saved in excel. At this data retrieval or crawling stage, take advantage of the libraries available in the python programming language with several parameters needed, namely keywords, language, and limits. The keyword parameter serves to find the desired keyword so that the tweet or post that enters only contains the word in the keyword, while the language parameter serves to set the desired language, and the limit parameter is the maximum limit of data you want to retrieve.

B. SCRAPING

The technique of automatically receiving information based on websites without having to copy it manually is called web scraping. Web scraping aims to search for specific information and then collect it on the new web. Web scraping only focuses on obtaining data using retrieval and extraction methods to make searching for something with varying data sizes easier [14].

C. LABELING

Labeling means that determining the class is based on a tweet or post whose work is done manually using labeling hoaxes & not hoaxes. Labeling is done manually by seeing many things that must be considered in putting a label on a tweet or post (FIGURE 2). At this stage, the labeling process on the training data is carried out manually. To determine whether to label a hoax from a tweet or post, you can use the help of official websites, such as covid19 sites, WHO sites, and other trusted websites, such as tempo fact checks or turnbackhoax. Each data will be searched for truth by using trusted websites such as fact checks (cekfakta.com/), tempo (www.tempo.co/), covid Indonesia (covid19.go.id/), and others.



id	username	ibuat_pad	tweet	LABEL	link	text bersih	sumber
1.48E+18	suara_ber	Mon Jan 1	India Berikan Vak	TRUE	https://www.tribunnews	india ikan	TWITTER
1.47E+18	readers_id	Mon Dec (	Indonesia Beli 1,9	TRUE	https://nasional.tempo.c	indonesia	TWITTER
1.48E+18	Beritasatu	Thu Jan 0(	Indonesia harus d	TRUE	https://www.beritasatu.c	indonesia	TWITTER
1.48E+18	suara_ber	Sun Jan 0(	Indonesia Kedata	TRUE	https://nasional.kompas.	indonesia	TWITTER
1.46E+18	Beritasatu	Thu Nov 1	Indonesia kembal	TRUE	https://www.beritasatu.c	indonesia	TWITTER

FIGURE 2. Twitter Labelling

D. PREPROCESSING

The process of reshaping unstructured text data as a structured form according to its needs can be called preprocessing [18]. The built system has five preprocessing stages, as shown in FIGURE 3 (TABLE 1).

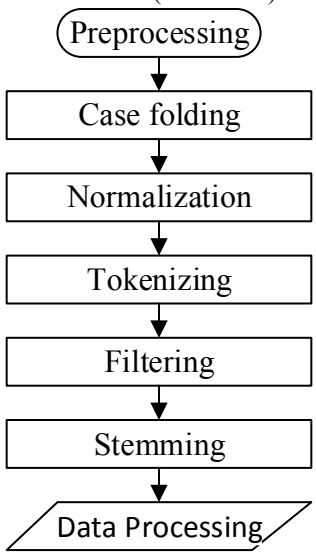


FIGURE 3. Data Preprocessing

1) CASE FOLDING

Case Folding is the process of changing all letters in the text to lowercase to make them uniform. Characters other than the letters a-z are omitted and considered delimiters.

2) NORMALIZATION/CLEANING

Normalization is converting all abbreviated words in the text into the corresponding word in the Language dictionary. Cleaning is the process of removing URLs, mentions, hashtags, punctuation marks, and numbers [17].

3) TOKENIZING

Tokenizing is a method to separate text or paragraphs into words.

4) FILTERING

Filtering is selecting essential words in the text of tokenizing results using a stop list algorithm (removing terms that are considered unimportant) or a wordlist algorithm (storing basic terms) [19].

5) STEMMING
It stems from updating the word that is affixed as a word that takes the form of an essential expression.

TABLE 1
Preprocessing

ORIGINAL DATA	PROCESSED DATA
Pfizer CEO's Statement on The Limitations Of The Efficacy Of Two Doses Of Covid-19 Vaccine	real pfizer ceo on two-dose effective limit on covid Vaccine
Robert Malone's Statement Regarding mRNA Vaccines for Covid-19 Have Not Been Adequately Tested and Children's Vaccinations Are Not Beneficial	real robert malone hook vaccine mrna covid test how on vaccination of children is not beneficial
Qatari Footballer Suffers Heart Attack after Being Vaccinated against Covid-19	Qatari football natural heart attack has been vaccinated against COVID
Pfizer and WHO Work Together to Bring Up Omicron Covid-19 Variant as Punishment for South Africa	pfizer who cooperates emerge variants of covid omicron like south african law
Pfizer Adds Substances Used to Stabilize Heart Attack Victims to Covid-19 Vaccine	pfizer adds substance to stabilize heart attack victims in covid Vaccine
Henry Kissinger's Speech Says Vaccines are a Tool for Population Control	speech henry kissinger vaccine control tool population
Jokowi's Regime to Hold Forced and Deadly Mass Vaccination in February 2022	jokowi regime holds mass vaccination of forced death in February
Thousands of People in Indonesia Died after The Covid-19 Vaccine	thousands of indonesians living in the world have been vaccinated against COVID

E. TERM PRESENCE
This process aims to calculate the frequency of the presence of a word on a document in calculating T.F (TABLE 2). using one type of formula, namely Binary T.F. Binary. Binary T.F. is helpful for paying attention to the occurrence of a word by giving a value of 1 for existing data. The data will be labeled as 0 if it does not exist. Furthermore, each word occurrence will be calculated by giving a 1 (there is) or 0 (none) (TABLE 3).

TABLE 2
Term Presence

DATA	TWEET	LABEL
data 1	reach | million | injectable | Indonesia | sign in | great | vaccine covid | world	true
data 2	indonesia | healthy | | vaccine covid	true
data 3	know | hoax | vaccine covid	fake
data 4	who | can | |information | post check out | fact	fake

TABLE 3
Term Calculation

NO	TERM	TRUE		FAKE	
		TWEET 1	TWEET 2	TWEET 3	TWEET 4
1.	tired	1	0	0	0
2.	million	1	0	0	0
3.	injection	1	0	0	0
4.	indonesia	1	1	0	0
5.	enter	1	0	0	0
6.	big	1	0	0	0
7.	vaccine	1	1	1	0
8.	covid	1	1	1	0
9.	world	1	0	0	0
10.	healthy	0	1	0	0
11.	know	0	0	1	0
12.	hoax	0	0	1	0
13.	who	0	0	0	1
14.	get	0	0	0	1
15.	information	0	0	0	1
16.	post	0	0	0	1
17.	refer	0	0	0	1
18.	facts	0	0	0	1

F. BAG OF WORDS
A bag of words is a process to calculate the number of values 1 and 0 on each label [22] (TABLE 4). The process has been estimated previously in the term presence. The calculation of 1 and 0 will then be used for the information-gaining process (TABLE 5).

TABEL 4
Bag of words value of 1

VALUE 1		LABEL		TOTAL
NO	TERM	TRUE	FAKE	
1.	tired	1	0	1
2.	million	1	0	1
3.	injection	1	0	1

18.	facts	0	1	1
Total		13	10	23

TABLE 5
Bag of words value of 0

VALUE 0		LABEL		TOTAL
NO	TERM	TRUE	FAKE	
1.	tired	1	2	3
2.	million	1	2	3
3.	injection	1	2	3

18.	facts	2	1	3
Total		23	26	49

G. INFORMATION GAIN
Information Gain is also defined as Mutual Information. Information Gain is a technique to reduce the number of appropriate or relevant features, then reduce the dimensions of the elements in the data to be used [5][20]. In equation 1, I.G. (A) is the difference in entropy values which is then stored by studying the variable A.

$$IG(A) = H(S) - \sum i \frac{Si}{S} H(S_i) \quad (1)$$

Where H(S) is an entropy-based on a given data set & A is an entropy based on the subset obtained using partitioning S according to feature A. In machine learning, Information Gain can be used to determine the order of features. Usually, features use a high Information Gain value and must be given a higher ranking than other features because they have a more vigorous intensity in classifying data [6] (TABLE 6).

TABLE 6
Information Gain Results

NUMBER	WORDS	INFORMATION GAIN
1.	covid	0.004413005
2.	vaccine	0.004297887
3.	vaccination	0.001910197
4.	booster	0.000496175
5.	not	0.000424789
6.	healthy	0.00035677
7.	child	0.000353549

1995.	fertile	5.03426E-07

H. NAÏVE BAYES MULTINOMIAL

Multinomial Naive Bayes is a probability learning method widely used in Natural Language Processing (NLP). Naive Bayes Multinomial includes developments based on the Naive Bayes method designed to handle text documents using word counts to be the underlying method of calculating probabilities. This method considers the number of words in the document. This method does not consider the context of the news and the order of the words in the document. Here's the equation of the formula in the Naïve Bayes Multinomial method Eq. (2) [7]:

$$P(c|document\ term\ d) = P(c) \times P(t_1|c) \times \dots \times P(t_n|c) \quad (2)$$

Where, P_c is the prior probability of class c; t_n is said document to – n; $P(c|term\ document\ d)$ indicates the probability of a copy belonging to class c; $P(t_{n|c})$ is the probability of the word to – n with known class c.

To select the prior probability value of class c can be calculated using the formula Eq. (3)

$$P(c) = \frac{N_c}{N} \quad (3)$$

where, N_c is the number of classes c on the whole document. N is the sum of the entire document.

The probability of the word to – n in class c can be determined using the Laplacian smoothing technique as follows Eq. (4):

$$P(t_n|c) = \frac{count(t_n c)+1}{count(c)+|V|} \quad (4)$$

Where, count $(t_n c)$ is the number of tn terms found throughout the rehearsal data with category c and coupled with a value of 1 to avoid a value of 0. $count(c)$ is number of terms across the trained data with category c. V indicates sum of all terms on the trainer data (TABLE 7).

TABLE 7
Naïve Bayes Multinomial Results

NO	TEXT	ACTUAL	PREDICT
1.	screenshot news rk invite abu janda prone to test covid Vaccine for china	FAKE	FAKE
2.	ask about coronavirus korlap fpi only the high priest against coronavirus high priest grandson of the prophet	FAKE	FAKE
3.	high rates check pcr fixed large rp thousand island java bal rp thousand island java bal	TRUE	FAKE

144.	issued international covid vaccine certificate health ministerial standards read full read	TRUE	TRUE

I. CONFUSION MATRIX

The Confusion Matrix is a stage of analysis & assessment of the performance of the designed system [15]. each row represents examples in an actual class and each column represents instances in a predicted class, are documented in the literature [16]. Performance is measured using the following calculations:

1) ACCURACY

The Confusion Matrix is a stage of analysis & assessment of the performance of the designed system. Performance is measured using the following calculations Eq. (5) [13]:

$$accuration = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

2) PRECISION

Precision is the level of accuracy between the information requested and the answers given by the system. The following is the formula for calculating the precision Eq. (6) [13]:

$$precision = \frac{TP}{TP + FP} \quad (6)$$

3) RECALL

A recall is a total of users correctly classified in a class divided using the total users in that class. The following is the formula for calculating recall Eq. (7) [13]:

$$recall = \frac{TP}{TP+FN} \quad (7)$$

III. RESULT

This research was conducted based on information about the covid19 Vaccine circulating on social media that prevents the government from stopping the spread of covid19 due to false news or hoaxes. Hoax news about the composition of the covid19 Vaccine, hoax news about the impact of the covid19 Vaccine, and lies about rejecting the covid19 Vaccine. The first thing to do is to prepare data from Twitter crawling, scraping from the website turnbackhoax.id, and taking manual data on the communication and information technology website used for testing. Data collection is carried out with keywords #vaksincovid19 with a total of 720 datasets labeled genuine or fake manually. The data that has been collected is processed using preprocessing text. Preprocessing is the initial process of preparing datasets into clean data that will be further processed. This study has five stages of preprocessing: Case Folding, Normalization, Tokenizing, Filtering, and Stemming. After the preprocessing stage is completed, the processing is done by selecting the information gain feature and the multinomial naïve Bayes method to get a genuine or fake value.

Researchers have also carried out the process of splitting data into training data and testing data with a ratio of 80:20 [23]. The training data was divided into several k-folds in this study into ten folds [24]. The accuracy value is the degree of proximity between the predicted and actual values. We use confusion matrix calculations to get the accuracy value score. A precision value is the degree of accuracy between the information requested in the answer given by the system and the recall value specified in the extent of the system's success in rediscovering knowledge. Researchers have carried out the confusion matrix testing process using 720 datasets divided as train data & the test data received an average accuracy value of 81.39%, precision of 80.36%, and recall of 79.73%. The highest accuracy is using k-fold two. The accuracy value reaches 88.8%, the precision value is 79.1%, and the recall value is 86.3%. The lowest accuracy was obtained on the 8th k-fold with an accuracy value of 73.6%, a precision of 75.4%, and a recall of 86.9%.

IV. DISCUSSION

Based on the results above, the accuracy value that has been obtained can answer the question "What percentage of tweets that are true are predicted to be hoaxes or those that are not hoaxes from all tweets?" In this case the accuracy value is 81.39% so it can be said that the result of the calculation is accurate. Precision describes the degree of accuracy between the requested data and the predicted results provided by the model. Thus, precision is the ratio of positive true predictions compared to the overall positive predicted results [21]. All the positive classes that have been predicted correctly; how much data is positive. From the calculations above, it answers the question of "What percentage of tweets are true hoaxes from the entire tweet that is predicted to be a hoax?". In this case the precision value is 80.3 6% which means that the model built is already quite precise. Recall describes the model's success in reinventing information. Thus, recall is the ratio of positive correct predictions compared to the overall positive true data [25]. From this study, calculating the recall value can answer the question "What percentage of tweets are predicted to be hoaxes compared to the entire tweet that is actually a hoax?". In this case the recall value is 86.87 6%. The results developed from this study produce more accurate values when compared to previous studies that have been carried out [2][3][7][11][12].

V. CONCLUSION

Based on the discussion and the results, the following conclusions can be drawn as follow; this study successfully used the multinomial naïve Bayes method and the weighting of the word information gain to detect Indonesian hoax content about the covid-19 Vaccine circulating on social media Twitter. The user acceptance test gives us an average score of 86.87%. Researchers have carried out the testing process with 720 datasets divided into training data and testing data, getting an average accuracy value of 81.39%, precision of 80.36%, and recall of 79.73%. This study succeeded in using data by crawling data from Twitter, scraping data from turnbackhoax.id websites, and taking data manually on the Communication and Information Technology website with a total of 720 datasets. The future works are stated as follow; in this study, it is only taking data and processing data from twitter, it is hoped that in future studies it can use more other social media. The calculation results not only display percentages but can be added with diagrams. In the crawling section, it is better to add a form to fill in the labelling manual.

REFERENCES

[1] C. Destitus, W. Wella, and S. Suryasari, "Support Vector Machine VS Information Gain: Analisis Sentimen Cyberbullying di Twitter Indonesia," *Ultim. InfoSys J. Ilmu Sist. Inf.*, vol. 11, no. 2, pp. 107–111, Dec. 2020, doi: 10.31937/si.v11i2.1740.

[2] F. Rahutomo, I. Y. R. Pratiwi, and D. M. Ramadhani, "Eksperimen Naïve Bayes Pada Deteksi Berita Hoax Berbahasa Indonesia," *J. Penelit. Komun. DAN OPINI PUBLIK*, vol. 23, no. 1, Jul. 2019, doi: 10.33299/jpkop.23.1.1805.

[3] K. M. Lhaksmana, F. Nhita, and A. Budhiarto, "Klasifikasi Pengguna Media Sosial Twitter Dalam Persebaran Hoax Menggunakan Metode Backpropagation Classification of Users Social Media Twitter in the Hoax Spread," *e-Proceeding Eng.*, vol. 4, no. 2, pp. 3082–3090, 2017.

[4] A. Prasetyo, Rino, Indriati, and P. Adikara, Pandu, "Klasifikasi Hoax Pada Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Modified K-Nearest Neighbor," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 12, pp. 7466–7473, 2018.

[5] D. N. Fitriana and Y. Sibaroni, "Klasifikasi Data Tweet dengan Menggunakan Metode Klasifikasi Multi-Class Support Vector Machine (SVM) (Studi Kasus : PT.KAI)," *e-Proceeding Eng.*, vol. 7, no. 2, pp. 8493–8505, 2020, [Online]. Available:

https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/12746.
[6] A. Wanto, "Optimasi Prediksi Dengan Algoritma Backpropagation Dan Conjugate Gradient Beale-Powell Restarts," *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 3, pp. 370–380, Jan. 2018, doi: 10.25077/TEKNOSI.v3i3.2017.370-380.
[7] B. S. Prakoso, D. Rosiyadi, H. S. Utama, and D. Aridarma, "Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifier Dengan Seleksi Fitur Dan Boosting," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 2, pp. 227–232, Aug. 2019, doi: 10.29207/resti.v3i2.1042.
[8] C. S. Sriyano and E. B. Setiawan, "Pendeteksian Berita Hoax Menggunakan Naive Bayes Multinomial Pada Twitter dengan Fitur Pembobotan TF-IDF," *e-Proceeding Eng. Vol.8, No.2*, vol. 8, no. 2, pp. 3396–3405, 2021.
[9] P. Gemilang and Y. A. Sari, "Klasifikasi Kategori Buku Ilmu Agama Islam Menggunakan Metode Naive Bayes Dan Seleksi Fitur Information Gain," vol. 4, no. 6, pp. 1650–1658, 2020.
[10] A. Rahman and A. Doewes, "Online News Classification Using Multinomial Naive Bayes," *ITSMART J. Ilm. Teknol. dan Inf.*, vol. 6, no. 1, pp. 32–38, 2017.
[11] A. I. AY Ananta, AP Kirana, "Implementasi Naive Bayes dan Pos Tagging menggunakan Metode Hidden Markov Model Viterbi pada Analisa Sentimen Terhadap Akun Twitter Presiden Joko Widodo Di Saat Pandemi COVID - 19," *Semin. Inform. Apl. POLINEMA*, no. ISSN 2460-1160, pp. 235–241, 2020.
[12] A. P. Kirana and A. Bhawiyuga, "Coronavirus (COVID-19) Pandemic in Indonesia: Cases Overview and Daily Data Time Series using Naïve Forecast Method," Indones. J. Electron. Electromed. Eng. Med. informatics, vol. 3, no. 1, pp. 1–8, Feb. 2021, doi: 10.35882/ijeeemi.v3i1.1.
[13] Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (PDF). Journal of Machine Learning Technologies. 2 (1): 37–63. Archived from the original (PDF) on 2019-11-14.
[14] B. Zhao, "Web Scraping," in Encyclopedia of Big Data, Cham: Springer International Publishing, 2017, pp. 1–3.
[15] Stehman, Stephen V. (1997). "Selecting and interpreting measures of thematic classification accuracy". Remote Sensing of Environment. 62 (1): 77–89. Bibcode:1997RSEnv..62...77S. doi:10.1016/S0034-4257(97)00083-7.
[16] Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies. 2 (1): 37–63. S2CID 55767944
[17] A. Esuli and F. Sebastiani, "Training Data Cleaning for Text Classification," 2009, pp. 29–41.
[18] "Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data". Tableau. Retrieved 2021-10-17.
[19] Sheerin, Jude (29 March 2010). "How spam filters dictated Canadian magazine's fate". BBC Online. Retrieved 5 April 2011.
[20] Larose, Daniel T. (2014). Discovering Knowledge in Data: An Introduction to Data Mining. Hoboken, New Jersey: Wiley. pp. 174–179. ISBN 9780470908747.
[21] Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies. 2 (1): 37–63.
[22] Sivic, Josef (April 2009). "Efficient visual search of videos cast as text retrieval" (PDF). IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 31, NO. 4. IEEE. pp. 591–605.
[23] V. R. Joseph, "Optimal ratio for data splitting," Stat. Anal. Data Min. ASA Data Sci. J., vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/sam.11583.
[24] S. Purushotham and B. K. Tripathy, "Evaluation of Classifier Models Using Stratified Tenfold Cross Validation Techniques," 2012, pp. 680–690.
[25] S. A. Alvarez, An exact analytical relation among recall, precision, and classification accuracy in information retrieval. Chestnut Hill, MA 02467 USA, 2002.

ANNISA PUSPA KIRANA is a lecturer at the State Polytechnic of Malang, Indonesia. She holds a Master's degree in Computer Science from Bogor Agricultural University, Indonesia. Her research specialization is in data mining and artificial intelligence. Her research areas are satellite data processing, spatial and temporal data analysis, and resilience systems related to environmental disasters and pandemics. She is a recipient of different national and international research awards, such as DIPA and SEAMEO. Annisa has filed several intellectual property rights on her innovative ideas, such as a forest fire prediction system, e-learning system, and supervisor-examiner selection system using artificial intelligence. She has authored or co-authored more than 20 publications, with 4 H-index and more than 50 citations. She can be contacted at email: puspakirana@polinema.ac.id

GUNAWAN BUDI PRASETYO is a lecturer at State Polytechnic of Malang, Indonesia. He holds a doctoral degree from the University of Southampton, U.K. His research area is Business intelligence, data mining, and artificial intelligence. Currently, he is actively teaching in the Department of Information Technology. He can be contacted at email: gunawan.budi@polinema.ac.id

ELA WIDYA LESTARI is an Information Department student at State Polytechnic of Malang, Indonesia. Her main research is about text mining, especially in the field of social media. She can be contacted at email: ellawidya91@gmail.com.