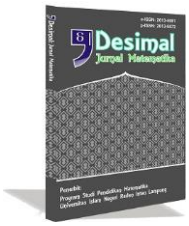
Desimal: Jurnal Matematika Vol 4 No 2 (2021) 193-202

Contents lists available at DJM

DESIMAL: JURNAL MATEMATIKA

p-ISSN: 2613-9073 (print), e-ISSN: 2613-9081 (online), DOI 10.24042/djm

http://ejournal.radenintan.ac.id/index.php/desimal/index



Sentiment analysis using fuzzy naïve bayes classifier on covid-19

Zhurwahayati Putri[1], Sugiyarto[1,*], Salafudin[2]

[1] Ahmad Dahlan University, Indonesia
[2] IAIN Pekalongan, Indonesia

ARTICLE INFO

Article History
Received : 01-11-2020
Revised : 25-07-2021
Accepted : 28-07-2021
Published : 30-07-2021

Keywords:
covid-19; Fuzzy Membership function; Fuzzy Naive Bayes Classiifier; Sentiment Analysis.

*Correspondence: E-mail:
sugiyarto@math.uad.ac.id

Doi:
10.24042/djm.v4i2.7390

ABSTRACT

Fuzzy Naive Bayes Classifier method has been widely applied for classification. The Fuzzy Naive Bayes method which consists of a combination of two methods including fuzzy logic and Naive Bayes is used to create a new system that is expected to be better. This research aim to find out the society's sentiments about COVID-19 in Indonesia and the use of the results of the Fuzzy Naive Bayes Classifier. The data of this research is obtained by scraping on Twitter in the period from January 1, 2020 to April 30, 2020. The classification method used in this research is the Fuzzy Naive Bayes Classifier method by applying the fuzzy membership function. In this research, sentiment analysis uses input data whose source is taken from tweets and the output data consists of sentiment data which is classified into three classes, namely positive class, negative class, and neutral class. In the distribution of training and testing data of 70%: 30%, the accuracy of the classification model using the confusion matrix is 83.1% based on 1199 tweet data consisting of 360 testing data and 839 training data. Also the presentation of each sentiment class was obtained which was dominated by positive sentiments, namely the positive class by 36.7%, the negative class by 35.0%, and the neutral class by 28.3%. Based on the results of the presentation, it can be concluded that there are still many people who have positive opinions or give positive responses to the presence of COVID-19 in Indonesia.

http://ejournal.radenintan.ac.id/index.php/desimal/index

INTRODUCTION

The Fuzzy Naïve Bayes Classifier Method is a modification method that sets two systems, namely fuzzy logic and Naive Bayes. The aim is to make a new system that is expected to be better. Fuzzy logic is a method used to solve problems that are stored or have ambiguity. Fuzzy logic discusses the fuzzy set theory. This theory states that there are only 0 and 1 (not members of the set and members of the set), but in the range between 0 and 1 (Fathurochman et al., 2014). In systems that require complexity, fuzzy logic tends to be difficult and requires a long time to determine the appropriate membership

settings and rules. While Naive Bayes Classifier is a statistical classification method that can predict the probability of a class that supports Bayes theorem. Bayes theorem is a theory used to calculate the class opportunities of each group of variables (Buani, 2016). Naive Bayes invokes simple freedom, the relationship between variables in the class is entirely independent of other variables. This simple assumption results in suboptimal estimates of the expected results are inaccurate, specifically for text domain problems (Zaidi et al., 2013). Tree and K-NN, have also found that accuracy and speed are the factors that most support and utilize data in classifying data (Slamet et al., 2018). By taking advantage of the two methods, Fuzzy Naive Bayes Classifier is expected to be able to solve the problem of ambiguity in predicting the acquisition problem of a class.

Sentiment analysis is the study of opinions as well as the analysis expressed in the text (Liu, 2012), and a research problem that requires multiple sub-tasks of NLP, such as text extraction (Poria et al., 2016). Sentiment analysis is part of Natural Language Processing (NLP) which is useful for clarifying and classifying sentiment classes in the text form. Classes are classified as positive, negative, and neutral, but we can discuss classes into three categories which are not easy because they cause ambiguity in a text that makes it difficult to classify. Some other problems also exist in the very positive and very negative sentiment classes or only positive and negative factors taken into account (Balahur, 2013). Sentiment analysis can be further evaluated using the Naive Bayes classification where the data from social media.

Social media is a means used by people to join one another by creating, sharing, and exchanging information and creating virtual networks and communications (Nasrullah, 2015). Some examples of social media are Facebook, Twitter, Instagram, Line, WhatsApp, and others. Social media fixing this feature is constantly being calculated from the move. Judging from the role and benefits of social media has an important role for us, one of which provides information.

The information provided can take a variety of forms, such as issues related to health, politics, business, country, and so on. The world is currently in an uproar with a plague attacking humanity. This outbreak is a new type of disease called novel coronavirus or n-coronavirus. This novel coronavirus was officially announced as a pathogen causing COVID-19 on January 8, 2020, by the Center for Disease Control and Prevention of China (Li et al., 2020). The epidemic of COVID-19 began in Wuhan, China last December and has become a significant public health problem, not only in China but also in countries around the world (Phelan et al., 2020). The World Health Organization (WHO) announces that this outbreak is an emergency health problem of international concern (Mahase, 2020). Judging from the amount of information circulating about COVID-19, this is where social media can take advantage of its role as an information service provider for the community. Many online media that contain information related to COVID-19 such as Twitter.

Twitter can be described as a microblog or social network. This site is for microblogging because its main activity is posting short status update messages (tweets) through the web. Tweets are short messages that are limited to 140 characters. Because of these characteristics, many microblogging services (quick and short messages) make spelling mistakes, using emoticons and other characters that express certain meanings on Twitter (Agarwal et al., 2011). This media always provides up-to-date information from time to time. People often talk about and give their opinions or opinions on topics given through

information from the media. The forms of opinions and opinions given through the online media are varied, some support (agree), and some that conflict with what is presented by each of the online media.

Author further researched the opinions or opinions given by the community, then sentiment will be collected from Twitter based on the information presented. So from a sentiment, it can provide information that will be useful. The researcher wants to know the public sentiment in the presence of COVID-19 in Indonesia and the use and results of the Fuzzy Naive Bayes Classifier in COVID-19 sentiment analysis.

METHOD

The Naive Bayes Classifier is a probability-based machine learning technique. The Naive Bayes Classifier is a simple method but has high accuracy and performance in text classification. The Naive Bayes Classifier has two main processes, namely the training and testing processes. The training process will produce a classification rule that contains the probability of each variable (conditional probability) and the target class of the data training that has been determined. The testing process is a process of testing each variable to determine the optimal class based on the classification rules that have been built (Fathurochman et al., 2014).

In Naive Bayes Classifier, each tweet is notified as a variable pair $(x_1, x_2, x_3, \dots, x_n)$ where x_1 is the first word, x_2 is the second word, x_3 is the third word, and so on. Then there is V which is defined as a class set. The classification stage in the Naive Bayes Classifier is done after the preprocessing stage so that the data used at this stage is clean. There are four processes carried out at the classification stage, namely (Nugroho et al., 2016):

1. Formation of Training Data Variables

This process is the initial process in the classification stage of the Naïve Bayes Classifier, where the variable in question is a tweet (comment) from Twitter which will be classified into a sentiment class, namely positive, negative, or neutral, by paying attention to the frequency of appearance of each word in the variable.

2. Calculating Probability Data of Training Data

After forming the training data variable by paying attention to the frequency of occurrence of each word in the variable, the next process is to calculate the probability of each class using the following equation:

$$P(v_j) = \frac{n(v_j)}{n(S)} \qquad (1)$$

3. Determine Class Probabilities

In this process, the probability of each world in each class can be calculated using the following question:

$$P(x_I|v_j) = \frac{n_i + 1}{n + f_k} \qquad (2)$$

$n(v_j)$ is the number of classes in the category j , where j denotes the order of the data being tested, and $n(S)$ is the number of classes used in the training process.

4. V_{MAP} Calculation

In classification, the Naive Bayes Classifier method will produce the class with the highest probability (V_{MAP}) of the testing data in each class by entering variable pairs $(x_1, x_2, x_3, \dots, x_n)$. The following is (V_{MAP}) presented in following equation.

$$(V_{MAP}) = \arg\max_{v_j} P(v_j) \prod_{i=1}^{n} f(x_i|v_j) \qquad (3)$$

The classic set (crisp) is usually defined as a collection of elements or objects $x \in X$, X finite and countable. Every single element may or may not be a

set A, where $\subseteq X$. The statement "x is a member of A" can be true, while in other cases this statement can be false. Such classical sets can be described in different ways, such as defining member elements using a characteristic function, where 1 denotes membership and 0 is non-membership (Zimmerman, 1996). In this research, a set element does not only consist of 1 and 0, but it is in the range of 1 to 0 as in the definition of the following fuzzy set.

Definition (Zimmerman, 1996) If X is a collection of objects that are generically denoted by x, then a fuzzy set $\tilde{A}$ in X is a set of sequential pairs:

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x)) | x \epsilon X\}$$

$\mu_{\tilde{A}}(x)$ is called the membership function or degree of membership (also the degree of compatibility or degree of truth) of x in $\tilde{A}$ which maps x to the membership space of M which is located in the range [0,1].

Membership function is a curve that shows the mapping of data input points into the value of membership (often also called the degree of membership) which has an interval between 0 to 1 (Zimmerman, 1996). Several functions can be used, including the triangle curve as follows:

The triangle curve is a combination of two lines (linear) as in Figure 1.

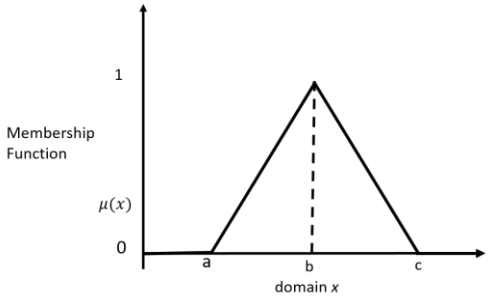


Figure 1. Triangle curve

Membership Function:

$$\mu(x) = \begin{cases} 0; & x \leq a \; atau \; x \geq c \\ \dfrac{(x-a)}{(b-a)}; & a < x \leq b \\ \dfrac{(c-x)}{(c-b)}; & b < x < c \end{cases} \quad (4)$$

The confusion matrix is a matrix that provides information on the comparison of the classification results performed by the system (model) with the actual classification results. The confusion matrix is in the form of a matrix table that describes the performance of the classification model on a series of testing data whose actual values are known.

Generalizations for confusion matrices with more than two or n classes are presented in Table 1 below:

Table 1. Model Evaluation

		Predicted Class			
		Class 1	***Class 2***	***...***	***Class n***
True Class	*Class 1*	x_{11}	x_{12}	...	x_{1n}
	Class 2	x_{21}	x_{22}	...	x_{2n}
	...	...	...	...	...
	Class n	x_{n1}	x_{n2}	...	x_{nn}

From table 1 above, a confusion matrix can be formed as follows:

$$Confusion\ Matrix = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nn} \end{bmatrix}$$

Based on the confusion matrix above, the accuracy with more than two or n classes is (Manliguez, 2016):

$$Accuracy = \frac{\sum_{j=1}^{j=n} x_{i,j}}{number\ of\ testing\ data} \quad (5)$$

RESULTS AND DISCUSSION

Distribution Data and Preprocessing Stages

This research uses two data consisting of input data and output data. Input data consists of tweet data, and the output data consists of sentiment data

which has been classified into three classes, namely positive, negative, and neutral. Input and output data are obtained by crawling on Twitter. The distribution of training data and testing data in this research was 80%: 20%; 70%: 30%; 60%: 40%; and 50%: 50%. The distribution of training data and testing data used in the calculation simulation is 70%: 30%.

The next step is the preprocessing process in which tweets are replaced with lowercase letters, replacing the words, and finally, the words are filtered using a lexicon dictionary. The lexicon dictionary contains unimportant words. The training data and testing data before and after the preprocessing stage of data obtained from January 1, 2020, to April 30, 2020, are:

Table 2. Training Data

No.	Tweet	Tweet Preprocessing	Sentiment
1	Dunia Bingung Hingga Saat Ini Belum Ada Kasus Virus Corona Di Indonesia http://dlvr.it/RPgrfs pic.twitter.com /UhwAj3cvJc	*dunia bingung hingga saat ini belum ada kasus virus corona di indonesia*	neutral
2	Bantu Pencegahan Virus Corona di Indonesia, Nikita Mirzani Sumbang Rp 100 Juta, Ini Harapannya https://medan.tribunnews.com/20 20/03/18/bantu-pencegahan-virus-corona-di-indonesia-nikita-mirzani-sumbang-rp-100-juta-ini-harapannya … via @tribunmedan	*bantu cegah virus corona indonesia nikita mirzani sumbang rp 100 juta ini harap via*	positive
3	Setelah di serang virus Corona. Indonesia menyusul dilanda Virus Kegoblokan. https://twitter.com/BadjaNuswantara/status/1234820717151014913 …	*telah serang virus corona indonesia susul landa virus goblok*	negative
4	Kaget itu sdh menjadi wabah di Indonesia, sama seperti Corona di Cina	*kaget sdh jadi wabah indonesia sama corona cina*	negative
5	Melonjak 893, Ini Data Sebaran Kasus Positif Virus Corona di Indonesia https://correcto.id/beranda/read/22124/melonjak-893-ini-data-sebaran-kasus-positif-virus-corona-di-indonesia …	*lonjak 893 ini data sebar kasus positif virus corona indonesia*	positive
6	Peneliti Harvard: Virus Corona Seharusnya Sudah Ada di Indonesia https://www.suara.com/tekno/2020/02/09/071500/peneliti-harvard-virus-corona-seharusnya-sudah-ada-di-indonesia?utm_source=twitter&utm_medium=share …	*teliti harvard virus corona harus sudah ada indonesia*	neutral

Table 3. Testing Data

No	Tweet	Tweet Preprocessing	Sentiment
1	Belum Ada Kasus Virus Corona di Indonesia - https://www.starjogja.com/2020/01/22/belum-ada-kasus-virus-corona-di-indonesia/ …pic.twitter.com/n1OKdWhSu0	*belum ada kasus virus corona indonesia*	neutral
2	Line Sediakan Update Kasus Corona di Indonesia https://www.solopos.com/line-sediakan-update-kasus-corona-di-indonesia-1052435?...	*line sedia update kasus corona indonesia*	positive
3	Gara2 virus corona udah di Indonesia, masuk pabrik aja karyawannya harus di scan. Udah kayak produk swalayan aja cuk wkwk	*gara2 virus corona udah indonesia masuk pabrik aja karyawannya di scan udah kayak produk swalayan aja cuk wkwk*	negative

Stages of Naive Bayes Classifier

Based on the discussion in Naïve Bayes Classifier, the following are the results of the Naive Bayes Classifier stage:

1. Formation of Training Data Variables

This process is the initial process in the Naive Bayes Classifier classification stage, where the variable in question is a tweet (comment) from Twitter that will be classified in sentiment class that is positive, negative, or neutral, taking into account the frequency of occurrence of each word in the variable. The results of the formation of training data variables are presented in the table below.

Table 4. The Results of the Formation of Training Data Variables

No	Sentimen	100	893	aja	bantu	bingung	cina	...	wabah	wkwk
1	*neutral*	0	0	0	0	1	0	...	0	0
2	*positive*	1	0	0	1	0	0	...	0	0
3	*negative*	0	0	0	0	0	0	...	0	0
4	*negative*	0	0	0	0	0	1	...	1	0
5	*positive*	0	1	0	0	0	0	...	0	0
6	*neutral*	0	0	0	0	0	0	...	0	0

2. Calculating Probability Data of Training Data

In the previous training data sharing process, it was known that each tweet has a sentiment that has been classified into three sentiment classes, namely positive, negative, and neutral classes, then the probability of each sentiment class will be calculated using equation (1). The probability results of each class in the training data are presented in the table below:

Table 5. Frequency Probability Results from Appearance

v_j	$n(v_j)$	$P(v_j)$
negative	2	0,33333
neutral	2	0,33333
positive	2	0,33333

3. Determine Class Probabilities

In the process of forming the training data variable, seen from the frequency of word occurrences in each sentiment class, the number of words that appeared in the positive class was 21 times, in the negative class 14 times, and in the neutral class 10 times and the total of all words in the training data are 45 words. Thus, based on equation (2), the results of the probability of each word in each sentiment class are as follows:

Table 6. The Results of the Probability of each Word in each Sentiment Class

v_j	kata							
	100	**893**	**aja**	**bantu**	**bingung**	**...**	**wabah**	**wkwk**
negative	0,01695	0,01695	0,01695	0,01695	0,01695	...	0,03390	0,01695
neutral	0,01818	0,01818	0,01818	0,01818	0,03636	...	0,01818	0,01818
positive	0,03030	0,03030	0,01515	0,03030	0,01515	...	0,01515	0,01515

4. V_{MAP} Calculation

The V_{MAP} calculation on the testing data is carried out to find the highest probability in each sentiment class, the highest probability value is the sentiment class of the tweet on the testing data based on the values obtained in the training data. The following is the formation of testing data variables that will be used to obtain V_{MAP}, where the process of forming testing data variables is the same as the process of forming the training data.

Table 7. The Results of the Formation of Testing Data Variables

No	Sentimen	100	893	aja	bantu	bingung	cina	...	wabah	Wkwk
1	*neutral*	0	0	0	0	0	0	...	0	0
2	*positive*	0	0	0	0	0	0	...	0	0
3	*negative*	0	0	2	0	0	0	...	0	1

Based on the results of the formation of the testing data variables from Table 7, it is found that the words that are in the positive class, seen from the frequency of their appearance, are in the form of the words "corona, Indonesia, line, provide, update", words that are in the negative class are in the form of the word "aja, corona, cuk, gara2, Indonesia, etc".

Table 8. The Results of V_{MAP} Calculation from Testing Data

No	*negative*	*neutral*	*positive*
1	0,0000438	0,0000541	0,0000313
2	4,20E-09	5,96E-09	2,40E-09
3	2,46E-26	7,06E-26	4,58E-27

Implementation of the Fuzzy Membership Function

The corresponding curve in this research is a triangle curve. The process starts with determining the values of x, a, b, and c. The value of x will be taken from the V value of MAP in each sentiment class, while the value of $a = 0, b = 0.5$, and $c = 1$. From the calculation using a triangle curve, the most dominant value of each of the existing V_{MAP} classes will be taken, so that the results can be presented as follows:

Table 9. The Results of Fuzzy Membership Function Calculation from Testing Data

No	*negative*	*neutral*	*positive*	*F_negative*	*F_neutral*	*F_positive*
1	0,0000438	0,0000541	0,0000313	0,0000876	0,000108	0,0000626
2	4,20E-09	5,96E-09	2,40E-09	8,40E-09	1,19E-09	4,80E-10
3	2,46E-26	7,06E-26	4,58E-27	4,92E-27	1,41E-26	9,16E-28

The confusion matrix of the Naive Bayes Classifier is as follows:

$$Confusion\ Matrix = \begin{bmatrix} 112 & 8 & 6 \\ 6 & 87 & 9 \\ 22 & 10 & 100 \end{bmatrix}$$

Based on the confusion matrix above, the accuracy of the classification model in this sentiment analysis can be calculated by equation (5), namely:

$$Accuracy = \frac{112 + 87 + 100}{112 + 8 + 6 + 6 + 87 + 9 + 22 + 10 + 100} = \frac{299}{360} = 0.831$$

with the presentation of each sentiment class namely positive class at 36.7%, negative class at 35.0%, and neutral class by 28.3%.

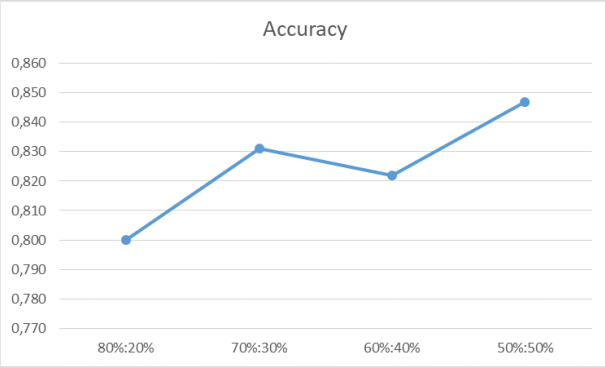


Figure 2. Accuracy of Comparison of Training Data and Testing Data

CONCLUSIONS AND SUGGESTIONS

This research produces a sentiment analysis on COVID-19 using the Fuzzy Naive Bayes Classifier method. In the research process using tweets by doing Scrapping on Twitter using Python 3.7 software as input data. The tweets were classified into three sentiment classes, namely positive, negative, and neutral. The results of testing on the distribution of training and testing data of 70%: 30% obtained an accuracy of 83.1% based on 1199 tweet data consisting of 360 testing data and 839 training data. The percentage of each sentiment class is obtained, namely the positive class of 36.7%, the negative class of 35.0%, and the neutral class of 28.3%. Based on the results of the presentation, it can be concluded that there are still many people who have positive opinions or give positive responses to the presence of COVID-19 in Indonesia.

REFERENCES

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). *Sentiment analysis of twitter data*. Department of Computer Science. Columbia University.

Balahur, A. (2013). *Sentiment analysis in social media texts*. European Commission Joint Research Centre.

Buani, D. C. P. (2016). Optimasi algoritma naïve bayes dengan menggunakan algoritma genetika untuk prediksi kesuburan (fertility). *Program Studi Teknik Informatika, STMIK Nusa Mandiri Jakarta*, *4*(1).

Fathurochman, D., Witanti, W., & Yuniarti,

R. (2014). Perancangan game turn based strategy menggunakan logika fuzzy dan naive bayes classifier. *Conference: Seminar Nasional Informatika At: Yogyakarta*, *1*(1).

Li, Q., Guan, X., Wu, P., Xiaoye, W., & et al. (2020). Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia. *N Engl J Med [Epub Ahead of Print 29 Jan 2020] in Press*. https://doi.org/10.1056/NEJMoa2001316

Liu, B. (2012). *Sentiment analysis and opinion mining: Synthesis lectures on human language technologies*. Morgan and Claypool Publishers. https://doi.org/https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Mahase, E. (2020). China coronavirus: WHO declares international emergency as death toll exceeds 200. *BMJ.368:M408*. https://doi.org/10.1136/bmj.m408

Manliguez, C. (2016). Generalized confusion matrix for multiple classes. *Cinmayii Manliguez. University of the Philippines*. https://doi.org/10.13140/RG.2.2.31150.51523

Nasrullah, R. (2015). *Media sosial; persfektif komunikasi, budaya, dan sosioteknologi*. Simbiosa Rekatama Media.

Nugroho, Didik, G., & Yulison Herry Chrisnanto, A. W. (2016). Analisis sentimen pada jasa ojek online menggunakan metode naïve bayes. *Program Studi Informatika, Fakultas Matematika Dan Ilmu Pengetahuan Alam, Universitas Jenderal Achmad Yani, Semarang*.

Phelan, A. L., Katz, R., & Gostin, L. O. (2020). The novel coronavirus originating in wuhan, china. *China: Challenges for Global Health Governance [Epub Ahead of Print 30 Jan 2020] in Press. JAMA*. https://doi.org/10.1001/jama.2020.1097

Poria, S., Cambria, E., & Gelbukh, A. (2016). *Aspect extraction for opinion mining with a deep convolutional neural network. Knowledge Based Systems*. 42–49.

Slamet, C., Andrian, R., Maylawati, D. S., Suhendar, Darmalaksana, W., & Ramadhani, M. A. (2018). Web scraping and naive bayes classification for job search engine. *IOP Conf. Ser.: Mater. Sci. Eng. 288 012038.* https://doi.org/10.1088/1757-899X/288/1/012038

Zaidi, N. A., Cerquides, J., Carman, M. J., & Webb, G. I. (2013). Alleviating naive bayes attribute independence assumption by attribute weighting. *Journal of Machine Learning Research*, *14*, 1947–1988.

Zimmerman, H. J. (1996). *Fuzzy sets theory and its applications*. Massachusetts: Kluwer Academic Publishers.