Sentiment Analysis of Pancasila Values in Social Media Life Using the Naive Bayes Algorithm

1st Kusnawi Kusnawi
Informatics, Faculty of Computer Science
University of AMIKOM Yogyakarta
Indonesia
Email: khusnawi@amikom.ac.id

2nd Abiyoga Hendra Wijaya
Informatics, Faculty of Computer Science
University of AMIKOM Yogyakarta
Indonesia
Email: abiyoga.wijaya@students.amikom.ac.id

Abstract - The implementation of Pancasila in citizenship from year to year has not found good news, even in mid-2020 there were many conflicts due to polemics carried out by the Pancasila Ideology Development Agency in the bill. Various efforts have been made by both state officials and activists to make people aware of the importance of implementing Pancasila values in social life. Therefore, this study aims to determine whether a person has Pancasila spirit or not in terms of that person's social media. The research will result in the accuracy of Naive Bayes classification using the media in the form of tweets reaching 73% with the results in the form of 3 classes, namely Neutral, Positive and Negative Pancasila. The results of this study have resulted and can be applied to classify someone with Pancasila spirit or not by using media in the form of social media accounts with the Naive Bayes method, being able to detect early as a basis for decision making by stakeholders and can trigger someone to evaluate himself, apply Pancasila values in every activity, even on social media.

Keywords: Pancasila, Naive Bayes, Sentiment Analyst, social media

I. INTRODUCTION

It has been more than 75 years that Pancasila has been formulated and then used as the basis for the establishment of the Unitary State of the Republic of Indonesia in the implementation of the life of the state and the nation has yet to find a bright spot. In fact, from year to year, criminal acts committed by individuals who have positions in Indonesia do not decrease but increase. This is evidenced by the recent fact that many Indonesian citizens have been subject to criminal acts, which are mostly caused by the misuse of social media. In the use of social media, humans are indeed made easier to communicate, share information and explore the world without having to lose a lot of time and money, but social media can also be a place to spit hatred, insults, pornography, and even be used as a place to increase existence without using restrictions.[1]

Indonesia is one of the countries that upholds the ethics of good manners in communicating and behaving, but if it is associated with the digital world which has a universal concept in establishing communication, hyper-connectivity, speed in accessing information makes users trapped in the virtual world to forget the code of ethics in communicating on social networks becomes a serious problem when it is associated with the current situation where all activities are carried out virtually. [2]

The government's policy to carry out large-scale social restrictions has forced Indonesian citizens to change all forms of daily activities online through social networks and this has become one of the gaps for citizens to abuse and use social media as a platform for committing crimes. what if this is not considered properly it will result in the decline of the Indonesian millennial generation in terms of etiquette and manners and the further distance of the Pancasila spirit in a person and this will have an impact on the quality of human resources in Indonesia.

So, it is necessary to pay attention to take early preventive action by making ethical corrections. When they do not have good ethics or good personalities, they need direction to improve human resources from an early age. even training programs to build good ethics can be one way to prevent corruption. This is because relying on a universal industrial code of ethics cannot cover all situations faced by society due to different cases of position [3]. Therefore, it is necessary to detect ethical ownership and professionalism in a person by classifying that person's Pancasila (Pancasila) spirit.

The nature of Pancasila itself is a form of someone who has behaved based on Pancasila, in which Pancasila can be a basis for doing something to reflect good citizens of the country [4]. Therefore, there needs to be an innovation or breakthrough, a system that can detect how much a person has the chance to become a Pancasila or not, so we can take early precautions against actions that person can take someday.

Based on the background that has been presented, researchers are interested in detecting the spirit of Pancasila using the Naïve Bayes classification method (Pancasila Recognition Using Naïve Bayes classifier). A test of one of the classification methods (Naïve Bayes Classification) is one of the first steps that can be taken to analyze whether a person has the opportunity to have a Pancasila spirit or not. There are many ways to classify a person, which can be seen from the way of speaking, signatures, or sitting, as well as from handwriting. However, in this case, the researcher wants to try to classify using social media.

II. LITERATURE

Jasman Pardede, et al (2020), conducted a study to find out comments that contained bullying elements using the naïve Bayes method. The results of the research are in the form of an application that is used to input comments which will be processed later to determine whether the comment has cyberbullying elements or not [5].

Bayu Yudha Pratama and Riyanarto Sarno (2016), conducted a study to find the right method for conducting personality classification. This study resulted in the conclusion that classifying personality using the Naïve Bayes classifier method has higher accuracy than using KNN and SVM [6].

Muslim Nasyroh and Rinandita Wikansari (2017) conducted a study on whether personality has a relationship with one's performance. This study uses a quantitative approach with questionnaires distributed to several employees [7].

Tia Sari Indayani (2019), researched to classify a person's personality using the Support Vector Machine 7 (SVM) method. This study used 1500 data taken from Twitter by using a classification approach based on the Big Five Personality [8].

A. Text Mining

Text Mining is a method of retrieving large amounts of textual data which will be processed to produce information from the patterns that have been obtained [9]. Text mining can be said to be a method that is used to mine unstructured data which in the end will be analyzed to obtain information from that data [10].

B. Data Crawling

Data crawling is a process of retrieving data from a company that is carried out legally with the help of Application Programming Integration provided by the company[11]. This research will use a dataset in the form of tweet data taken from Twitter with the help of APIs that have been provided by Twitter. The following is the flow of data collection and storage of the dataset.

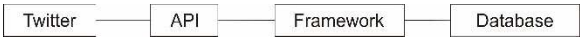

Fig. 1. Data Crawling

C. Preprocessing

When the text data has been obtained, the next step is to do preprocessing. Preprocessing is a process where the data that has been taken will be cleaned of words that have no meaning and also make them in the form of basic words so that the data becomes clean and ready to be processed. In Preprocessing, there are several stages of the process that will be carried out, namely Case Folding, Tokenizing, Stop word Removal, Stemming.

Case Folding : Change the data that has been taken into lowercase data

Tokenizing : Word-based cutting of data

Stop word Removal : The process to eliminate words that have no meaning and have no effect on the analysis process will be carried out later

Stemming : Returns the words that have been stored into the root word

D. Labeling

To produce training data and test data, the researcher equates the dataset that has been taken from Twitter and then equates it with the sentiment dictionary that has been prepared by the researcher. The framework can be described as follows:

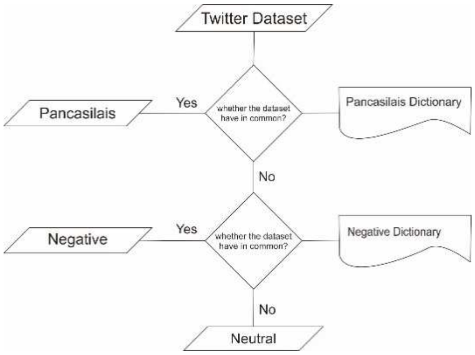

Fig. 2. Labeling

E. Naïve Bayes Classifier

Naïve Bayes is a classification method based on probability with the assumption of strong independence. The steps for classifying using the naïve Bayes method are counting the number of classes and then counting the number of data that have the same cases, after that step, continue to multiply the data that have similarities and compare them. The data that has the largest comparison results will be used as the final decision of the classification process using the naïve Bayes method [12].

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \qquad (1)$$

X : Class of unknown data

H : Hypothesis Data that already has a specific class value

P(H|X) : The probability depends on the condition X or can be called the posterior probability (data affected by the probability of the sample data)

P(H) : Data that has a true probability value from the sample data or called the prior probability

P(X|H) : The probability that depends on the condition H or can be called the likelihood

P(X) : Data Probability X or it can be called the Prior Probability predictor

The Naive Bayes method has advantages compared to other methods, including the use of relatively little training data to determine the estimated parameters that will be needed in the classification process. In this method, each document has an attribute denoted (x1, x2, x3, …xn).

xn is a symbol for the nth word in the document while the document class is symbolized by "V". So that in the classification process, the system will look for the highest likelihood or probability of the category symbolized "VMAP" [13].

The equation can be written as follows:

$$V_{map} = argmax_{vj\in v}\left(\frac{p(x_1,x_2,x_3,\ldots.x_n,|\,V_j)p(V_j)}{p(x_1,x_2,x_3,\ldots.x_n)}\right) \qquad (2)$$

P (x1, x2, x3, ... xn) has a constant value for all categories (Vj) so that it can be rewritten with the following equation:

$$V_{map} = argmax_{vj\in v} p(x_1, x_2, x_3, \ldots.x_n,|\,V_j)p(V_j) \qquad (3)$$

To make it easier to calculate the equation P (x1, x2, x3, ... xn). this equation can be simplified into

$$V_{map} = argmax_{vj\in v} \prod_{i=1}^{n}\left(p(x_1|\,V_j)p(V_j)\right) \qquad (4)$$

Below is an example of training data in the form of 3 tweets that have been processed beforehand, with each tweet having a different class and number of vocabulary. From the three tweets, the probability value of each class will be searched which will later be used to perform calculations on the tweet test data that will be classified.

TABLE 1. P(V_J) VALUE IN EACH CLASS

Data	Tweet data has been preprocessed	Sentiment	P(V_j)
1	["allah","rabi","sempurna","keluarga", "nikmat","sejahtera","lindung","dunia" ,"akhirat","amin"]	Pancasila	0,3
2	["ankara","guncang","bom","mobil","tewas"]	Negative	0,3
3	["ina","tim","teruji","swafoto","lomba ","irit"]	Neutral	0,3

When already know the value of of each class, the next step is to find the value of the probability of words in each class, as an example of searching for a probability value of the word Allah in each class using the formula:

$$p(x_1|\,V_j) = \frac{n_k+1}{n+|n\ vocabulary|} \qquad (5)$$

TABLE 2. THE PROBABILITY VALUE OF THE WORD ALLAH FOR EACH CLASS

V_j	Pancasila Sentiment		Negative Sentiment		Netral Sentiment	
	N_k	N	N_k	N	N_k	N
$P(X_i\,\mid V_j)$	1	9	0	5	0	6
	2 / 31		1 / 26		1 / 27	

The calculation is carried out repeatedly on each data that will be classified. For example, doing calculations with examples of tweet preprocessing data

["allah","guncang","keluarga","teruji","irit","sejahtera"]

The first step to performing calculations is calculating the probability value of each class

$$Vmap = V_j(Pancasila,Netral,Negative)\ P(V_j)пP(X_i|V_j) \quad (6)$$

TABLE 3. THE VMAP VALUE OF EACH CLASS

Word	Vmap (Pancasila)	Vmap (Negative)	Vmap (Neutral)
P(V_j)	0,3	0,3	0,3
Allah	2/31	1/26	1/27
Guncang	1/31	2/26	1/27
Keluarga	2/31	1/26	1/27
Teruji	1/31	1/26	2/27
Irit	1/31	1/26	2/27
Sejahtera	2/31	1/26	1/27
Total (*)	2,6996736 x 10^{-9}	1,83033238 x 10^{-9}	3,07887169 x 10^{-9}

From the calculation results, the value of sentiment in the neutral class is greater than the Pancasila class and negative so it can be concluded that the data set has a neutral class.

F. Confusion Matrix

After classifying, it is necessary to assess the level of accuracy, recall, and precision to determine the quality of the classification results. The Confusion Matrix method is one method that can be used to calculate all of them [14]. Confusion Matrix can use the table matrix approach as follows [15].

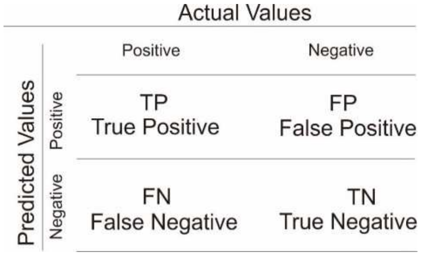

Fig. 1. Confusion Matrix [15]

$$Accuration = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \quad (7)$$

$$Precicion\ Probability = \frac{TP}{TP+FP} \times 100\% \quad (8)$$

$$Sensitivity\ Probabilitas = \frac{TP}{TP+FP} \times 100\% \quad (9)$$

III. METHODOLOGY

The data used in this study are in the form of tweets taken from someone's Twitter account. The per-account dataset has a quantity of the last 100 tweets created by the user. While the accounts selected as research data are in the form of 100 accounts which are divided into several groups, namely

TABLE 4. DATA ACCOUNT

Category	Artist	Placeman	Religious Figures	General Public
Number Of Accounts	16%	16%	16%	62%

This research is used to calculate how likely a person is to have a Pancasila spirit with several stages. These stages begin with collecting an account dataset which will later be used as training data and test data. Then the data that has been taken goes into the pre-processing process which consists of Case Folding to convert all data to lowercase. Then the data resulting from Case Folding is cut on a word-by-word basis using Tokenizing. The next stage is Stop word Removal or eliminating words that are deemed meaningless. The last process of pre-processing is to return words to root words using stemming. After going through pre-processing the data will be processed to get modelling to get results in the form of training data and test data. When the training data and test data are obtained, the test data will be processed using the naïve Bayes classification method with data samples in the form of available training data. After the test data is processed, you will get the final classification results which will be calculated for the accuracy of the data using the Confusion Matrix method. for more details will be illustrated in the following chart.

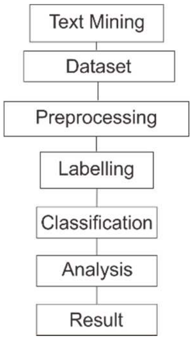

Fig. 4. General Structure

IV. IMPLEMENTATION AND DISCUSSION

A. Labelling

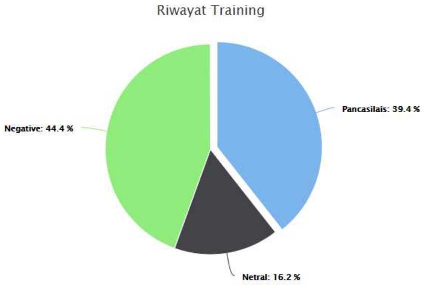

Fig. 5. Training Data

The labelling process resulted in training data with 16% labelled Pancasila 10% labelled negative 16.4% labelled neutral. The results of the labelling were taken 25% of the data and were classified using the Naive Bayes classifier.

B. Classification

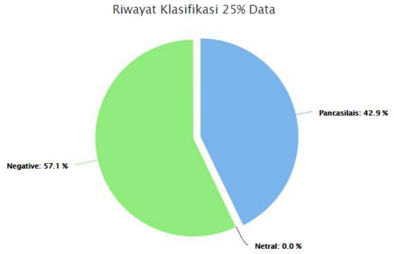

Fig. 6. Test Data

The classification process of 25% of the training data resulted in a percentage of 57,1% classified as Pancasila, 42,9% classified as negative, 0,0% classified as neutral.

V. ANALYSIS

From the results of the analysis of the test results data, 25% mixed data both Pancasila and negative with as many as 26 accounts to be analysed. From this data, the result is a calculation using the confusion matrix as follows:

TABLE 5. CONFUSION MATRIX

Matrix		Predicted Class		
		Pancasila	Negative	Neutral
Actual Class	Pancasila	647	152	1
	Negative	83	807	0
	Neutral	127	180	16

Based on the table above, the researchers got results in the form of 614 data on True Pancasila, 147 for False Pancasila, 80 for False Negatives, and 782 for True Negatives. After getting the prediction data for each class, the next step is to calculate the accuracy, precision, and recall.

$$Percentage\ Accuracy\ = \frac{1470}{2013} \times 100\ \% = 73\ \%$$

The results of the calculation of the percentage of accuracy above, the researcher can find out that the level of accuracy for classifying someone belonging to Pancasila by using the naïve Bayes method reaches 73%.

$$Precission\ = \frac{\sum \frac{True\ P}{True\ P + False\ P} \times 100\%}{l} \times 100\ \% \qquad (10)$$

a. Precision Pancasila class

$$Pancasila = \frac{647}{800} \times 100\ \% = 80\ \%$$

b. Precision Negative Class

$$Negative = \frac{807}{952} \times 100\ \% = 84\ \%$$

c. Precision Neutral Class

$$Netral = \frac{16}{17} \times 100\ \% = 94\ \%$$

d. Precision Al Class

$$Precission = \frac{80 + 84 + 94}{3} \times 100\ \% = 86\ \%$$

It can be concluded that the level of precision of classifying someone into Pancasila by using the naïve Bayes method can reach 86%.

$$Recall = \frac{\sum \frac{True\ P}{True\ P + False\ N}}{l} \times\ 100\% \qquad (11)$$

a. Recall Pancasila

$$Recall\ Pancasila = \frac{647}{647 + 153} = 0.80$$

b. Recall Negative

$$Recall\ Negative = \frac{807}{807 + 83} = 0.90$$

c. Recall Neutral

$$Recall\ \text{Netral} = \frac{16}{16 + 127} = 0.11$$

d. Recall All Class

$$Recall\ Semua\ data = \frac{1.81}{3} \times 100\% = 60\%$$

It can be concluded that the recall rate of classifying someone as Pancasila by using the naïve Bayes method can reach 600%.

TABLE 6. CALCULATION ACCURATION MATRIX

Matrix		Calculation		
		Accurate	Precision	Recall
Class	Pancasila	73%	80%	80%
	Negative		84%	90%
	Neutral		94%	11%
Average			86%	60%

VI. CONCLUSION

Based on the results of the analysis that has been carried out in chapter 6 using confusion matrix concludes:

a. This study using 100 account data with each account having a maximum tweet data of 100 tweets per account. By using test data totalling 25 accounts and analysis of test results using the confusion matrix method, it can be concluded that Naïve Bayes has an accuracy rate of 73% in performing classification.

b. By using the naïve Bayes classifier, researchers can classify based on data that has been taken from the Twitter account.

c. For now, the Naïve Bayes Method is concluded to be the right method for detecting whether a person has a Pancasila spirit or not.

d. From the results of the sentiment analysis of tweet data that has been carried out, it can be concluded that the results of the analysis can describe whether the community or a group of people has a Pancasila spirit or not.

e. From the results of the recall calculation, it can be concluded that the class that has the highest recall value is the negative class with a percentage of 90%, so that the prediction of test data with classes other than negative will be more likely to be classified as negative compared to changes to other classes

f. Resulting in the conclusion that the neutral class has a higher precision value than the other classes with a percentage of 94%. However, the recall results show that the recall value for the neutral class is only 11%. This is refraction caused by the system being able to only classify a small number of neutral classes

REFERENCE

[1] F. & A. A. Afriani, “Penerapan Etika Komunikasi di Media Sosial: Analisis Pada Grup WhatsApps Mahasiswa PPKn Tahun Masuk 2016 Fakultas Ilmu Sosial Universitas Negeri Padang,” *J. Civ. Educ.*, vol. 3, no. 3, pp. 331–338, 2020.

[2] Y. Fahrimal, “Netiquette: Etika Jejaring Sosial Generasi Milenial Dalam Media Sosial,” *J. Penelit. Pers dan Komun. Pembang.*, vol. 22, no. 1, pp. 69–78, 2018.

[3] M. Sohail and S. Cavill, “Accountability to Prevent Corruption,” no. September, pp. 729–738, 2008.

[4] P. Trijono, “Pancasila numerik,” 2016.

[5] J. Pardede, Y. Miftahuddin, and W. Kahar, “Deteksi Komentar Cyberbullying Pada Media Sosial Berbahasa Inggris Menggunakan Naïve Bayes Classification,” *J. Inform.*, vol. 7, no. 1, 2020.

[6] M. Pundlik Kalghatgi, M. Ramannavar, and N. S. Sidnal, “A Neural Network Approach to Personality Prediction based on the Big-Five Model,” *Int. J. Innov. Res. Adv. Eng.*, vol. 8, no. 2, pp. 2349–2163, 2015.

[7] M. NASYROH and R. Wikansari, “HUBUNGAN ANTARA KEPRIBADIAN (BIG FIVE PERSONALITY MODEL) DENGAN KINERJA KARYAWAN,” *J. Ecopsy*, vol. 4, no. 1, p. 10, May 2017.

[8] T. S. Indayani, “KLASIFIKASI KEPRIBADIAN BIG FIVE PERSONALITY BERDASARKAN TWEET MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM) TUGAS AKHIR,” 2019.

[9] J. N. Apriliana, Natalis Ransi, “Implementasi Text Mining Klasifikasi Skripsi Menggunakan Metode Naïve Bayes Classifier,” *Semant. Vol.3, No.2, Jul-Des 2017*, vol. 3, no. 2, pp. 187–194, 2017.

[10] L. P. Jing, H. K. Huang, and H. B. Shi, “Improved feature selection approach TFIDF in text mining,” *Proc. 2002 Int. Conf. Mach. Learn. Cybern.*, vol. 2, no. November, pp. 944–946, 2002.

[11] J. Eka Sembodo, E. Budi Setiawan, and Z. Abdurahman Baizal, “Data Crawling Otomatis pada Twitter,” no. August, pp. 11–16, 2016.

[12] M. F. Arifin, “Jurnal Inovasi Penelitian,” *J. Inov. Penelit.*, vol. 1, no. 3, pp. 1–4, 2020.

[13] E. F. U. Latifah, “Perbandingan Kinerja Machine Learning Berbasis Algoritma Support Vector Machine dan Naive Bayes (Studi Kasus: Data Tanggapan Mengenai Traveloka Melalui Media Sosial Twitter),” 2018.

[14] J. Han, M. Kamber, and J. Pei, “Third Edition : Data Mining Concepts and Techniques,” *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2012.

[15] H. Leidiyana, “Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor,” *J. Penelit. Ilmu Komputer, Syst. Embed. Log.*, vol. 1, no. 1, pp. 65–76, 2013.