International Journal of Artificial Intelligence & Robotics (IJAIR) E- ISSN: 2686-6269
Vol.1 No.1, 2019, pp.9-15

Personality Classification through Social Media Using Probabilistic Neural Network Algorithms

Mohammad Zoqi Sarwani[1]*, Dian Ahkam Sani[2], Fitria Chabsah Fakhrini[2]
Universitas Merdeka Pasuruan, Pasuruan, 67129, Indonesia
Email: [2]zoqi.sarwani@unmerpas.ac.id*; [2]dianahkam@unmerpas.ac.id; [3]ficha.fakhrini19@gmail.com
*corresponding author

ABSTRACT

Today the internet creates a new generation with modern culture that uses digital media. Social media is one of the popular digital media. Facebook is one of the social media that is quite liked by young people. They are accustomed to conveying their thoughts and expression through social media. Text mining analysis can be used to classify one's personality through social media with the probabilistic neural network algorithm. The text can be taken from the status that is on Facebook. In this study, there are three stages, namely text processing, weighting, and probabilistic neural networks for determining classification. Text processing consists of several processes, namely: tokenization, stopword, and steaming. The results of the text processing in the form of text are given a weight value to each word by using the Term Inverse Document Frequent (TF / IDF) method. In the final stage, the Probabilistic Neural Network Algorithm is used to classify personalities. This study uses 25 respondents, with 10 data as training data, and 15 data as testing data. The results of this study reached an accuracy of 60%.

Keywords: text mining; text processing; term inverse document frequent; probabilistic neural network.

I. INTRODUCTION

Today the internet is becoming a new digital space and producing a new generation, a generation raised in a modern cultural environment with interactive digital media and computer literacy. Old or traditional media began to be displaced by digital media, including social media like Facebook [1]. Indonesia probabilistic as the fourth-largest Facebook user after the USA, Brazil, and India. There are 65 million Facebook users, with 33 million users who open Facebook every day [2].

Social media is a site where anyone can create a personal page, to share information, and communicate with friends who are connected. They can get to know each other better, even though they have never met in person. According to Primada Qurrata Ayun, the emergence of social media makes someone merge her privacy space into public space. Next, he did not hesitate to show personal activities and show the mood to all friends through social media [3]. With this openness, a person's personality can be seen from the status of Facebook or expressions of mood on social media.

There are several methods or tests in psychology that can be used to determine a person's personality, including MBTI (Myers-Briggs Type Indicator), DISC (Dominance, Influence, Steadiness, Compliance), and Big Five. The Big Five personality is divided into five categories, namely Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N).

O personality is a personality that tends to actively imagine, sensitive to aesthetics, care about personal feelings, attracted to differences, intellectual curiosity, and freedom of opinion. C personality is a personality that tends to be able to control impulses, able to control themselves in careful planning, organization, and execution of tasks. E personality is a personality that tends to be active, has confidence, likes to talk even talkative, optimistic, likes to have fun and feels naturally cheerful. A personality is a personality that tends to put others first, has a sympathetic feeling towards others, and likes to help. N personality is a personality that tends to experience negative feelings such as feeling afraid, sad, awkward, angry, guilty, and hate [4].

Text mining is used to determine the source of knowledge in a document in the form of text. Text mining will produce data patterns, trends, and extraction of knowledge from potential text data. Text mining can also be referred to as data mining, where data in the form of text and data sources obtained from documents. The aim is to look for words that can represent the contents of the document. So it can be analyzed the relationship between documents. In general, the concept of the work of text mining is the same as data mining, namely predictive and descriptive excavation. However, text mining extracts a meaningful numerical index from the text, then the information contained in the text will be processed using various data mining algorithms.

Probabilistic Neural Network algorithm is an Artificial Neural Network (ANN) that can be used to solve classification problems. This algorithm is included in the stochastic algorithm, where variables are randomly determined and can change at any time by adjusting the situation. However, using the Probabilistic Neural Network algorithm the process can be carried out more quickly because the Probabilistic Neural Network algorithm requires only one training iteration [5].

This research is the development of a previous study entitled "Twitter Analysis to Know the Character of Someone Using the Naive Bayes Classifier Algorithm". The study analyzed a person's character through Twitter social media by using the MBTI

 DOI: 10.25139/ijair.v1i1.2025

(Myers-Briggs Type Indicator) test which is one of the methods in psychology to determine one's character [6]. To find out someone's character through social media that is twitter, the data used in analyzing is obtained from user tweets. The Naive Bayes Classifier algorithm is used to classify and the test results reach 100% accuracy when compared to the results of testing from experts. Furthermore, in the study titled "Personality Classification Based on Facebook Status Using the Backpropagation Method", states that information on a Facebook status that is considered not important and looks useless can apparently be used to find out a person's personality [7]. Someone's personality affects the work, so in this study, facebook status is used as data to find out a person's personality during the employee acceptance test. This method is expected to shorten the selection time of prospective employees. The algorithm used in this study is the Back-Propagation algorithm with a fairly high accuracy rate of 84.00%.

The research entitled "Campus Sentiment Analysis of E-Complaint Using Probabilistic Neural Network Algorithm", applies the Probabilistic Neural Network algorithm to classify student complaints through the website with the name e-complaint [8]. Student complaints are classified into two types, namely positive complaints and negative complaints. By using the Probabilistic Neural Network algorithm, the level of accuracy obtained is quite high, reaching 90%. A study with the title "Classification of Online News Using TF-IDF Weighting and Cosine Similarity", tf-IDF weighting and cosine similarity can be used to classify online news [9]. The data used comes from kompas.com. There is too much news posted on the web, so human errors often occur and do not fit into their categories. By weighting using tf-idf and cosine similarity processes can achieve the goal by being able to group news with an accuracy level of 91.25%.

II. METHODS

The flow of discussion in this paper is shown in Fig. 1, where the left side is the flow for training data and the right side is the flow for testing data. The training data flow begins with document retrieval on Facebook. Data were taken in the form of a facebook status document. The next step is in the form of text mining by doing text processing. At this stage, the document is processed to change the words in the status document to standard words.

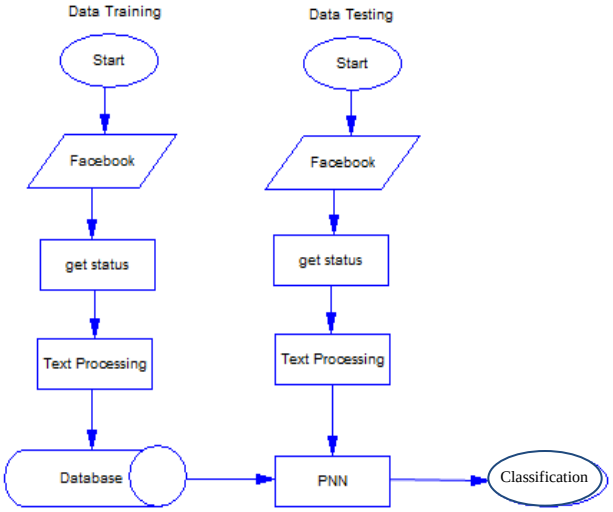

Fig.1. Research FLow

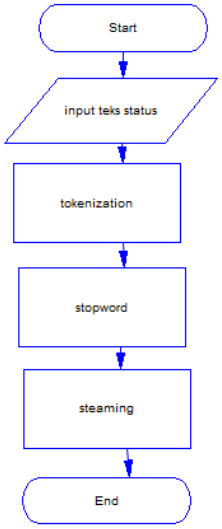

Fig.2. Flowchart Text Processing

The first step is to change the words on Facebook status into standard words, with the flow of the process can be seen in Fig. 2. After getting the standard words, then saved into the database as training data. Next, prepare the testing data by retrieving status documents on Facebook. This document is processed with text processing, to produce words that are standard and neater. The same process is done to get the testing data but does not need to be stored in a database. Testing data is calculated using the Probabilistic Neural Network algorithm to find out the appropriate classification. In the Probabilistic Neural Network algorithm, the first process to do is enter the input layer by entering a document in the form of status. The document is changed to the default word so that it can be processed at the layer. After that, every word is searched for the weight value. The weight value is used for the calculation of the next layer. To calculate the weight of each word, the TF / IDF method is used. The second process is the Pattern Layer. In this layer the calculation of each word in each category. Last is the Summation Layer, in the form of adding up words. In this layer the classification is determined based on the highest value.

Fig.2 is a plot of text processing, with the following process:

a) Tokenizing is a process to separate words that makeup sentences. The process of separating words using delimiter or space. Tokenization is done to eliminate punctuation and numbers that have no meaning. In addition to eliminating punctuation and numbers, the Tokenization process is also carried out the ToLowerCase process or the process of changing each letter in a document to all lowercase. For example, there is the phrase "Kegagalan itu seperti lipatan

 DOI: 10.25139/ijair.v1i1.2025

kertas yang pada waktunya akan menjadi origami yang indah". The sentence will be separated into tokens and change capital letters into lowercase letters so that the results are obtained, “kegagalan”, “itu”, “seperti”, “lipatan”, “kertas”, “yang”, “pada”, “waktunya”, “akan”, “menjadi”, “origami”, “yang”, dan “indah”.

b) Stopword is a word that often appears on documents. Stopwords are meaningless or connecting words. Stopword must be removed, because it can interfere with the classification process. Some examples of stopword "atau", "dan", kemudian", "yang", "kamu", and many more. Example in the sentence "Kegagalan itu seperti lipatan kertas yang pada waktunya akan menjadi origami yang indah. After tokenizing, the process will result in “kegagalan, “lipatan”, “kertas”, “waktunya”, “origami”, dan “indah”.
c) Steaming is the process of converting words into original or basic forms. This process parses words or deletes words that have prefixes, infixes, and postfixes so that they get basic words. Examples of affixes include "me", "meng", "an", "di" and so forth. In the previous stopword results in the form of “kegagalan, “lipatan”, “kertas”, “waktunya”, “origami”, dan “indah”. So in steaming the results obtained are the default words. The results are “gagal, “lipat”, “kertas”, “waktu”, “origami”, dan “indah”.

A. *Weighting*

In this study using the Term Frequency Inverse Document Frequency (TF / IDF) method for weighting. Weighting is the giving of value or calculation of the number of words in a document that is calculated with a certain weighting scheme. In a document each word has a different level of importance. Each word is given an indicator called term weight, (Zafikri, 2010). This method serves to find a representation of the value of each document in the database.

In the first step, calculate the TF or Term Frequency value. Term frequency is counting the same number of words in the database and words in the testing data. Each word is given a weighting of 1. The next step is to calculate the value of the IDF or Inverse Document Frequency, which aims to find out the total number of words that appear in the document. The weight of each term can be formulated as in equation (1).

$$W = TF \times IDF \quad (1)$$

Where the W variable is the weight value, the TF variable is the word value that appears, and the IDF variable is the number of words that appear in the document.

B. Probabilistic Neural Network

This research uses the Probabilistic Neural Network algorithm. Probabilistic Neural Network is an algorithm used to solve classification problems. The classification process is carried out in only one stage so that the process is slightly faster compared to other Neural Network derivative algorithms. In the Probabilistic Neural Network there are three layers, namely the input layer, pattern layer, and summation layer. In this study there are five classifications namely, Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The flow of the Probabilistic Neural Network algorithm with the five classifications can be seen in Fig. 3

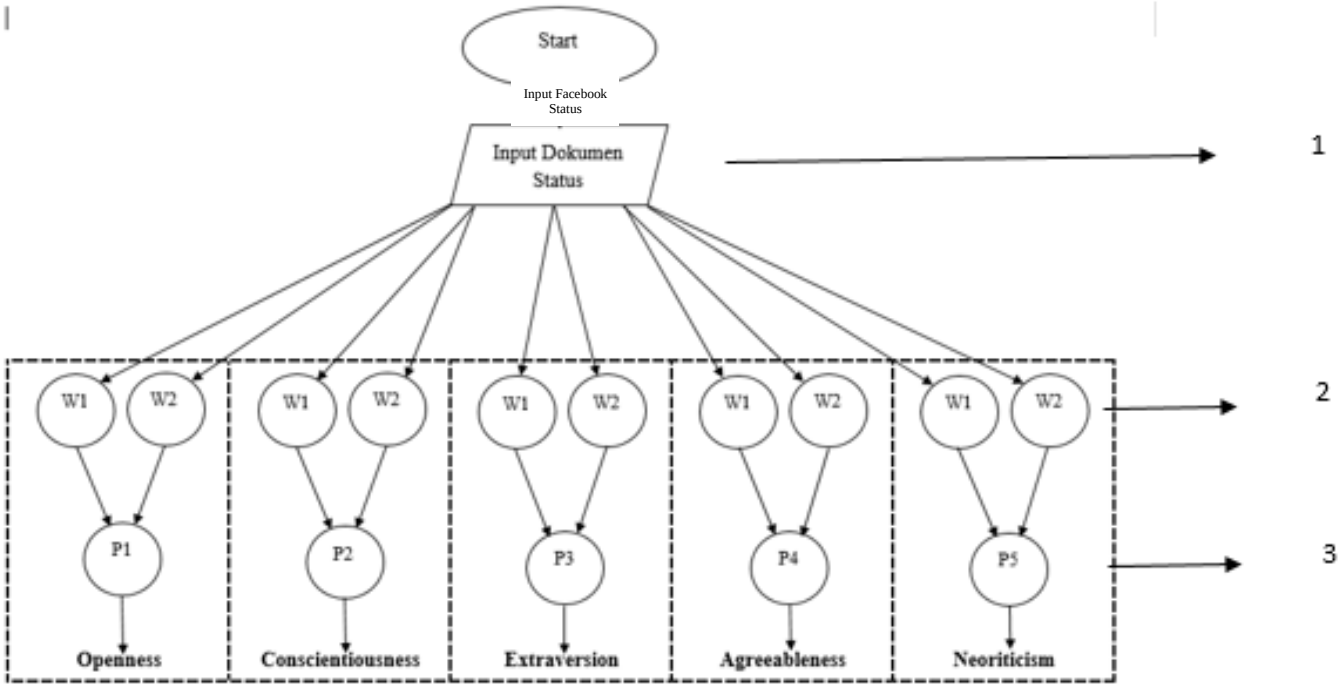

Fig.3. Probabilistic Neural Network Flow

 DOI: 10.25139/ijair.v1i1.2025

The Probabilistic Neural Network algorithm process starts from the input layer. In the drawing process, the input layer is symbolized by the number one which is located after the arrow that leads to right. The process at the input layer is to enter the document to be classified. The document contains the status of facebook users who have not been processed. The next process is text processing, to produce standard words or standard words that are considered to have meaning. The next process is the pattern layer, the pattern layer is located on the symbol of number 2. This process is calculated in each word. The variables W1 and W2 in the image are the weights of each word in the document. In the example there are two words exemplified so that there are only variables W1 and W2. The variable W is weight, while the numbers that follow behind it are constants of the number of filtered words.

Next is the value weighting process, to measure the most meaningful words or words that have the most influence on classification. To do the calculation, each word needs to be changed in numerical form. The goal is to convert words into numerics by weighting words. The weighting method used is the TF / IDF method. The calculation of TF / IDF can be seen in equation (1). The results of equation (1) are used as input in equation (2), where equation (2) is the pattern layer formula in the Probabilistic Neural Network algorithm. In the pattern layer each word is counted in each classification, so if the document contains three words, then the three numbers are counted in the five classifications. So each word goes through five calculations. If there are three words, then the document will go through five times the calculation. Whereas the pattern layer will be counted fifteen times because the pattern layer counts every word in each classification.

After getting the results on the pattern layer, the last process is to add all the results together. This process is included in the summation layer. The summation layer process is contained in the symbol number 3. In the picture there is the symbol P, this symbol is a symbol of the summation layer, where the number that follows is the number of classifications used. The summation layer also adds up all the values obtained using equation (3). The last process is the output or classification obtained. Classification is seen from the greatest value in the summation layer.

$$F_{ki}(d_j) = \frac{1}{2\pi\sigma^2} \exp\left(\frac{d_j^{\pi} w_{k,i^{-1}}}{\sigma^2}\right) \quad (2)$$

$$pi(d_j) = \sum_{k=1}^{N_i} F_{k,i}(d_j), i = 1,2,3,\dots c \quad (3)$$

C. Data Collecting

The data used in the study is facebook status. Before retrieving data, it is important to know the personality of the person being the target of the analysis. To find out someone's personality, then the first step taken is to spread the questionnaire as many as 25 questionnaires to active Facebook users. Next, the questionnaire data were analyzed by the experts concerned to find out his personality.

In the discussion of this paper, researchers conducted testing of research results in two stages. The first stage is testing the accuracy of text mining in processing data taken from Facebook social media. The second stage is testing the accuracy of the Probabilistic Neural Network (PNN) algorithm. In testing the data used as test material is the facebook user id. In this study facebook user id data were used as many as 25 facebook users, where 10 data were used as training data and 15 were used as testing data or training data.

D. Evaluation

Evaluation is carried out to determine the accuracy of an algorithm in this study. The accuracy of the results can be determined by comparing the results of the classification with the data from the results of expert analysis. Next these results are also compared with the total data or all data used both true and false data and multiplied by 100%. Calculation formula as in equation (4),

$$akurasi = \frac{\text{the correct amount of data}}{\text{the total amount of data}} x100\% \quad (4)$$

III. RESULT AND DISCUSSION

For starters, log in as admin by going to the top right button. Meanwhile, as a user, to start the system, you can directly enter the status you want to analyze his personality. Fig. 4 displays the number of personalities used for classification, i.e. there are five classifications, namely Openness, Conscientiousness, Extraversion, Agreeableness, and Neorithicsm. In this menu, the admin can add personality data, edit or delete. To add personality data by clicking the add button on the table, then a form will appear to fill in the new personality data. The personality data menu page is used to enter personality data to be used as a classification. The edit menu is used to edit data that has been entered when there are errors. The delete menu is used to delete data deemed unnecessary.

 DOI: 10.25139/ijair.v1i1.2025

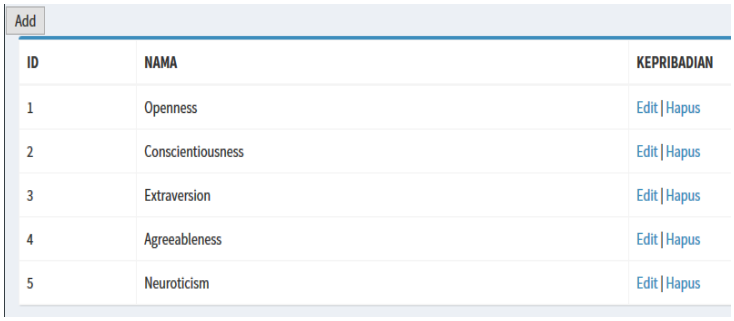

Fig. 4. Personality Data

The facebook data menu page in Fig.5 is used to view training data that contains the id, facebook name or facebook username, status, and id of personality type. To add data can be done by clicking the button located above the table and there is a form for filling new data. To edit or change data, there is an edit button on the right along with the delete menu.

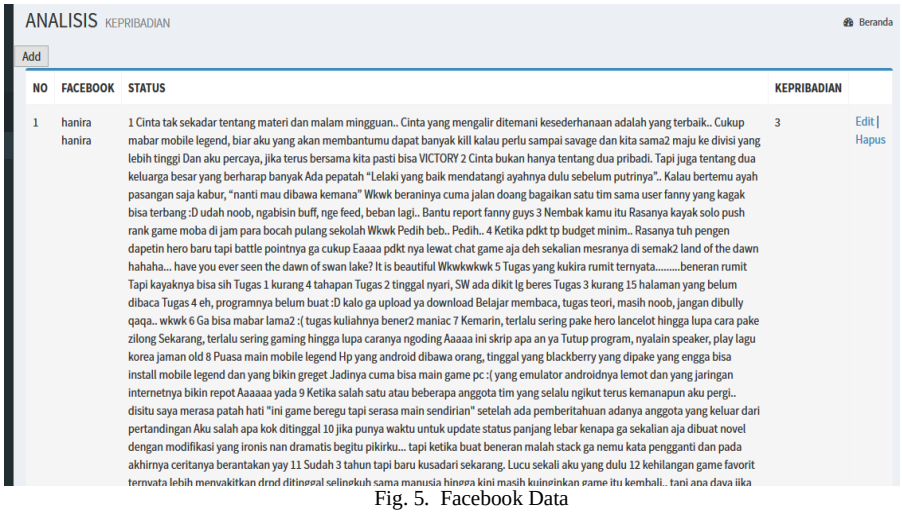

Fig. 5. Facebook Data

The Status Data page contains words from all the statuses in the training data that have been through the text processing process, wherein the process produced standard words or important words used to analyze a person's personality. As shown in Fig. 6

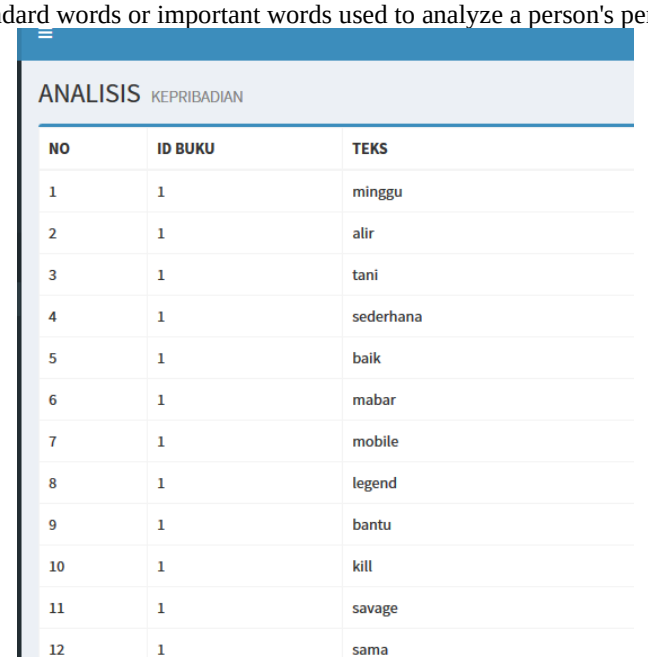

Fig. 6. Status Text

The results of personality tests carried out on 25 people with the test results as shown in table 1. In the data used to determine the form of personality.

DOI: 10.25139/ijair.v1i1.2025

TABLE I
RESULT OF ANALYSIS

Initials	Personality	Personality System	B/S
A	Extraversion	Extraversion	B
B	Neuroticism	Neuroticism	B
C	Agreeableness	Agreeableness	B
D	Agreeableness	Agreeableness	B
E	Openness	Openness	B
F	Neuroticism	Neuroticism	S
G	Neuroticism	Neuroticism	S
H	Conscientiousness	Conscientiousness	B
I	Openness	Openness	B
J	Openness	Openness	S
K	Extraversion	Extraversion	S
L	Agreeableness	Agreeableness	B
M	Agreeableness	Agreeableness	B
N	Extraversion	Extraversion	B
O	Extraversion	Openness	S
P	Openness	Openness	S
Q	Neuroticism	Neuroticism	S
R	Conscientiousness	Openness	S
S	Conscientiousness	Conscientiuosness	B
T	Openness	Openness	B
U	Extraversion	Extraversion	B
V	Agreeableness	Agreeableness	S
W	Neuroticism	Neuroticism	B
X	Conscientiousness	Conscientiuosness	B
Y	Conscientiousness	Conscientiousness	S

Based on the test results, it can be concluded that the Probabilistic Neural Network has an accuracy value using equation (4) of 60% with 15 users who are true or in accordance with personality, and 10 other users are false or not in accordance with their personality.

$$akurasi = \frac{15}{25} x100\% = 60\%$$

IV. CONCLUSION

Based on research from personality classification through social media using a probabilistic neural network algorithm using social media, one's personality can be analyzed, one of social media is Facebook. In classifying personalities through several processes including text processing and probabilistic neural network algorithms for classification. In the probabilistic neural network algorithm, there are three layers to analyze the status text, including the input layer, pattern layer, and summation layer. Text on the status will be entered via the input layer, then the frequency of each document is in the pattern layer, and the last layer or summation layer is a process for classification or classification results according to the document.

 DOI: 10.25139/ijair.v1i1.2025

ACKNOWLEDGMENTS

Furthermore, this research is expected to be developed more broadly by utilizing various social media that are now increasingly used by the community. Apart from that, it is hoped not only to process standard Indonesian but also slang. Other algorithm combinations are also possible for higher accuracy results.

REFERENCES

[1] I. S. Ibrahim, Kritik Budaya Komunikasi, Yogyakarta: Jalasutra, 2011.

[2] "Pengguna Internet di Indonesia 63 Juta Orang," 19 April 2019. [Online]. Available: http://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker.

[3] P. Q. Ayun, "Fenomena Remaja Menggunakan Media Sosial dalam Membentuk Identitas," *Channel*, pp. 1-16, 2015.

[4] A. T. Damanik and M. L. Khodra, "Prediksi Kepribadian Big 5 Pengguna Twitter dengan Support Vector Regression," *Cybermatika*, pp. vol. 3 - no. 1, 2015.

[5] E. Shofa, H. Yasin and R. Rahmawati, "Klasifikasi Data Berat Bayi Lahir Menggunakan Probabilistic Neural Network dan Regresi Logistik (Studi Kasus di Rumah Sakit Islam Sultan Agung Semarang Tahun 2014)," *Jurnal Gaussian*, pp. 815-824, 2015.

[6] M. Z. Sarwani and W. F. Mahmudy, "Analisis Twitter untuk mengetahui Karakter Seseorang Menggunakan Algoritma Naive Bayes Classifier," pp. 291-296, 2015.

[7] K. M. Lhaksamana, F. Nhita and D. Anggraini, "Klasifikasi Kepribadian Berdasarkan Status Facebook Menggunakan Metode BACKPROPAGATION," *e-Proceeding of Engineering*, p. 5174, 2017.

[8] M. Z. Sarwani and W. F. Mahmudy, "Campus Sentiment Analysis E-Complaint Using Probabilistic Neural Network Algorithm," *Kursor*, pp. 135 - 140, 2016.

[9] B. Heriwijayanti, D. E. Ratnawati and L. Muflikhah, "Klasifikasi Berita Online dengan Menggunakan Pembobotan TF-IDF dan Cosine Similarity," *Pengembangan Teknologi Informasi dan Ilmu Komputer*, pp. 306-312, 2018.

[10] A. Nilogiri, "Pengaruh Fitur Warna pada Klasifikasi Impresi Citra Batik Indonesia Menggunakan Probabilistic Neural Network," *JUSTINDO*, 2016.

[11] K. R. Prilianti and H. Wijaya, "Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering," *Cybermatika*, 2014.

[12] E. K. Putri and T. Setiadi, "Penerapan Text Mining pada Sistem Klasifikasi Email Spam Menggunakan Naive Bayes," *Jurnal Sarjana Teknik Informatika*, 2014.

 DOI: 10.25139/ijair.v1i1.2025