Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.6 No.1 January 2021, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Social Media Analysis Using Probabilistic Neural Network Algorithm to Know Personality Traits

Mohammad Zoqi Sarwani[1], Dian Ahkam Sani[2]

[1,2]*Informatics Engineering Department, Merdeka University Pasuruan, Indonesia*

[2]dianahkam@unmerpas.ac.id*

[1]zoqi.sarwani@unmerpas.ac.id

Abstract— The Internet creates a new space where people can interact and communicate efficiently. Social media is one type of media used to interact on the internet. Facebook and Twitter are one of the social media. Many people are not aware of bringing their personal life into the public. So that unconsciously provides information about his personality. Big Five personality is one type of personality assessment method and is used as a reference in this study. The data used is the social media status from both Facebook and Twitter. Status has been taken from 50 social media users. Each user is taken as a text status. The results of tests performed using the Probabilistic Neural Network algorithm obtained an average accuracy score of 86.99% during the training process and 83.66% at the time of testing with a total of 30 training data and 20 test data.

Keywords— Social Media, Probabilistic Neural Network, Personality Traits, Personality.

I. INTRODUCTION

Today, the internet is becoming a new digital space and producing a new generation. Generations raised in modern cultural environments or digital media are interactive and computer literate, and traditional media will change to digital media. One of them is social media, which is quite influential [1]. Some of the existing social media like Twitter, Facebook, path, etc. According to the Ministry of Communication and Informatics (*Kemenkominfo*), Indonesia has the fourth rank of Facebook users after the USA, Brazil, and India. There are 65 million active Facebook users, with 33 million active users per day [2].

Social media is a site where everyone can create a personal page, share information, and communicate with several friends connected with the user. Everyone involved in it feels that they know each other more than anything, even though they have never physically met face to face [3]. According to [2], social media's emergence makes a person dissolve their privacy space into public space. Making him do not hesitate to show his person's activities or moods by showing all his friends through social media [2]. With this openness, a person's personality can be seen from social media's status or mood expression.

Several methods or tests in psychology are used to determine personality, including the MBTI (Myers-Briggs Type Indicator), DISC (Dominance, Influence, Steadiness, Compliance), and the Big Five. The Big Five personality consists of openness (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N). The O personality is active in imagination, sensitive to aesthetics, cares about personal feelings, is interested in differences, intellectual curiosity, and freedom of opinion. C personality is closely related to impulse control, controlling oneself for careful planning, arrangement, and doing tasks. Personality E is an active person, has self-confidence, likes to talk or fussy, is optimistic, likes to have fun, and feels cheerful. Personality A always makes others first, has sympathy for others, and wants to help. The N personality tends to experience negative feelings such as fear, sadness, awkwardness, anger, guilt, and hatred [4].

Text mining is used to determine the source of knowledge in a document in text form. In the application using text mining, data patterns, trends, and extraction of knowledge from potential text data are obtained in previous research conducted by [5] and [6] used text mining to analyze sentiment. Also, [7] also use text mining for cases of character detection. The PNN (Probabilistic Neural Network Algorithm) algorithm is an Artificial Neural Network (ANN) that can be used to solve classification problems. By using PNN, the process can be carried out faster. This is because PNN only requires one training iteration [8]. [9], [10] and [11] also use the PNN algorithm to solve the case. Also, issues related to texts using the PNN method have been carried out by [6].

Based on the background above described by the researcher, this study uses a probabilistic neural network algorithm to find the model (features) and solve cases of detection of a personality based on the statistics contained in social media.

II. RESEARCH METHODOLOGY

Stages in this research, the author is inspired by and references previous research related to this journal's background problems. The research related to this journal includes: Twitter Analysis To Know A Person's Character Using The Naive Bayes Classifier Algorithm.

Research conducted by [7] analyzed a person's character through Twitter social media, character classification using the MBTI test, to determine a person's character through social media, using the Naive Bayes algorithm. Produce accurate results, where the classification carried out by experts and the classification using the Naive Bayes algorithm is the same.

DOI : 10.25139/inform.v6i1.3307

In the study [12], it was observed that Facebook status information could be used to determine a person's personality, especially for employee tests, making it easier and shortening the time for hiring employees. In this research, the algorithm used is Back Propagation. In his research, the accuracy rate obtained was 84.00%.

Selection of Smoothing Parameters on Probabilistic Neural Network Using Particle Swarm Optimization for Text Detect in Images Research conducted [13]. Detect text with little training data used the PNN algorithm. PNN algorithm to detect text has a high level of Accuracy 75.42% using only 300 data.

In the study [14][15], Facebook status information can be used to find out a person's personality, especially for student tests, making it easier and shortening the time for personality student. In [14] research, the algorithm used is PNN. The accuracy rate obtained was 60.00% from 25 respondents in his research, with 10 data training and 15 data testing. In [15], Naïve Bayes Classifier's algorithm with an accuracy of 88% from the same respondent.

Campus Sentiment Analysis E-Complaint Using Probabilistic Neural Network Algorithm In research [6], the PNN algorithm was applied to classify complaints in e-complaints. In his research, complaints are classified into two, namely positive complaints and negative complaints. By using the PNN algorithm, the accuracy rate reaches 90%.

In this study, several stages will be carried out. Namely, the first is data collection by taking a predetermined social media status. After that, perform the data preprocessing process. This process is carried out in several stages: case folding, tokenization, remove punctuation, stop words, standardization, and stemming. The next stage is weighting words using TF-IDF and performing the classification process using the Probabilistic Neural Network Method and the Evaluation process.

A. Data Collection

The data used in this study are social media status. The social media used in this study are only limited to Facebook and Twitter. The data collection process uses a scrapping technique and is stored in the form of CSV data. Social media users whose data were taken were students of *Universitas Merdeka Pasuruan* who had taken a psychological test. The status used in this study is only text status, not status in the form of images, videos, links, and icons. Also, the status taken is the latest status owned by the user. Table I show some example of data.

TABLE I
EXAMPLE DATA

User_a	*biar gak terlalu sumfek .. dangdutan aja dulu*
User_b	*Astaghfirullah, jauhkan saya dari orang ini ??biar gak ketawa terus*
User_c	*Cinta tak sekadar tentang materi dan malam mingguan.. Cinta yang mengalir ditemani kesederhanaan adalah yang terbaik.. Cukup mabar mobile legend, biar aku yang akan membantumu dapat banyak kill kalau perlu sampai savage dan kita sama2 maju ke divisi yang lebih tinggi :D*
User_d	*Bisa nggak sih kamu g seperti itu ke aq*

B. Pre-processing

Pre-processing is a step used to clean up data that is not sufficient to influence the classification process. This process is important because the data used at this stage is rough, so that the documents produced in this process can facilitate the classification process.

The preprocessing process was applied to all data used in this study. Furthermore, the data will be divided into 2 types of data: training data and testing data. The training data from the preprocessing results are used to form features, where this feature is used as a reference for carrying out the calculation process during training and testing. The preprocessing process has several stages that must be carried out, including:

1) Case Folding: Process used to change the capital letter on all social media statuses contained in the training data document and test data to lowercase

2) Tokenization: Process used to change each social media status's claims separated by each word as shown in Table II. This process applies to training data and testing data.

TABLE II
EXAMPLE OF TOKENIZATION

Sentence	*Bisa nggak sih kamu g seperti itu ke aq*
Tokenization	*"bisa","nggak","sih","kamu","g","seperti","itu","ke","aq"*

3) Remove Punctuation: Remove punctuation performs a process to remove all punctuation in the training data and test data.

4) Stopwords: After doing the tokenization process, the next step is to do the stopwords process. This process deletes individual words based on the word dictionary. The word dictionary is a list of conjunctions and words that are considered to have no influence or meaning in the classification process. Table III provides an example of the stopwords process.

TABLE III
EXAMPLE STOPWORD

Sentence	*Bisa nggak sih kamu g seperti itu ke aq*
Tokenization	*"bisa","nggak","sih","kamu","g","seperti","itu","ke","aq"*
Stopwords	*"bisa","nggak","kamu","g","seperti","aq"*

5) Standardization: Process standardization is a process used to convert abbreviated words into standard words. For example, the word "not" is changed to "no", the word "g" is changed to "no" and the word "*aq*" is changed to the word "I" as shown in Table IV.

DOI : 10.25139/inform.v6i1.3307

TABLE IV
EXAMPLES OF STANDARDIZATION

Stopwords	*"bisa","nggak","kamu","g","seperti","aq"*
Standardization	*"bisa","tidak","kamu","tidak","seperti","aku"*

6) Stemming: Process used to convert words into root words. This process is done by removing the affixes contained in each word. For example, the word "appreciate" will be changed to the root word "appreciate" by removing the affix "*meng*", the word "hit" is changed to the root word "hit" by removing the affix "*mem*".

C. TF-IDF Weighting

At the word weighting stage, the Term Frequency - Inverse Document Frequency (TF-IDF) method is used to get each word's weight value in each document. The word weighting process is carried out using the TF-IDF algorithm. TF-IDF works by counting the number of occurrences of a word multiplied by the log value of the number of documents used compared to the number of documents containing the word. The Tf-IDF method can provide maximum Value for unique words or words that do not frequently appear in other documents.

The formula used to calculate the TFIDF method is shown by equation (1).

$$X_i = TF_j \; x \; tdf_j \quad (1)$$

Where:
X_j = Value for the ith word
TF_i = The number of occurrences of the ith word
tdf_i =, where n is the number of documents$\log \frac{n}{df_i}$
df_i = number of documents containing the ith word

D. Probabilistic Neural Network (PNN)

A probabilistic Neural Network (PNN) is an algorithm that belongs to the neural network family. PNN has several layers and weights. The PNN algorithm performs the classification process with only one step so that it can provide faster output.

In this study, the PNN architecture used is 4 layers, including the input layer, pattern layer, summation layer, and output layer, as shown in Figure 1.

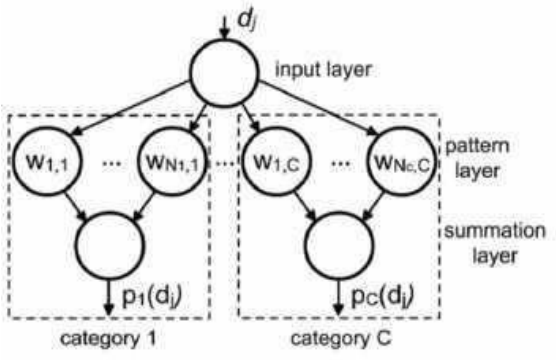

Figure 1. Architectures probabilistic neural network.

The PNN method performs the classification process by taking the largest probability value generated in the output layer. The stages used for the calculation process of the PNN method include:

- The first stage determines the PNN method's parameters in the form of the number of categories or classes and the number of neurons in the pattern layer and the Gaussian Value.
- Stage 2 performs a random weight generation process.
- Stage 3 performs the calculation process for each neuron in the pattern layer using equation (2) [9].

$$\phi(X, W_i) = e^{\frac{-(X-W_i)^T(X-W_i)}{2\delta^2}} \quad (2)$$

Where *X* variable is the input vector, Wi is the weight vector that connects the input layer and the pattern layer. And the parameter is a smoothing parameter.δ

- Stage 4 performs the calculation process on the summation layer by taking the neurons' sum in each class in the pattern layer. Equation (3) shows the formula for calculating the summation layer stage.

$$\sum_{k=1}^{n} \phi(X, W_i) \quad (3)$$

Where:
ϕ = Pattern layer at X, and Wi

- Stage 5 or the final stage carries out comparing each Value generated at the summation layer stage. Input layer The most considerable Value at the summation layer stage will be selected as the Value in the output layer and used to determine the selected class based on max values.

E. Testing

In this study, testing was carried out using various training data and tests to find the best classification model. Each combination of training data testing and test was run in 10 folds.

The accuracy value is obtained by comparing the number of correct data with the total number of data multiplied by 100%, as in equation (4). It is said that the data is correct if the results of the expert analysis are the same as the results issued by the system.

$$Accuracy = \frac{correct\ data}{total\ data} x100\% \quad (4)$$

III. RESULT AND DISCUSSION

Social media analysis uses a probabilistic neural network algorithm to find out someone, aiming to find out whether the status written by users on social media describes the character they have.

The data used for testing is the social media status data for Facebook and Twitter. Social media users whose status were taken were students of the Universitas Merdeka

DOI : 10.25139/inform.v6i1.3307

Pasuruan who had taken a psychologist test. Ten each was taken from the social media Facebook and Twitter. All user statuses are collected into 1 status document to get the user character results from the user status document.

Testing the combination of training data and testing data is done by dividing the amount of data. The data sharing technique was carried out by randomly selecting data in each class with a predetermined composition. The probabilistic neural network method is a stochastic algorithm. Testing is done by taking the average score of Accuracy obtained from the running process 10 times for each combination.

TABLE V
TESTING RESULTS OF THE COMBINATION OF TRAINING DATA AND TEST DATA

total			Average Accuracy (%)	
Training Data	**Features**	**Test Data**	**Testing**	**Testing**
15	4137	35	81,334	56,665
20	4593	30	85	71
25	5241	25	82	75.6
30	5312	20	86.99	83,666
35	5957	15	79.14	73,713

Table V shows that the combination of training data and test data with the best average score of Accuracy when the number of training data is 30 and test data is 20 with an accuracy score of 86.99% and 83.666%. From these results, we can see that when the training data is less than 30, the average accuracy score is more than 75%. But during the testing process, the Accuracy average score obtained was not satisfactory. So it can be said that the resulting model has not provided the best accuracy score. Meanwhile, when the training data is more than 30, it gives an average Accuracy score that is unsatisfactory during the training and testing process. This shows that the increasing number of features does not guarantee a model that can provide the best results because features can be increased and become noise.

IV. CONCLUSION

From the analysis and testing results to analyze social media using a probabilistic neural network algorithm to determine a person's character in the training and testing process, an average accuracy of 86.99% and 83.66% was obtained. Accuracy is obtained by a combination of training data and test data of 30 and 20, respectively. The test results show that the number of features cannot determine the best classification model. So that researchers suggest continuing this research by adding methods for feature selection.

REFERENCES

[1] Ibrahim, Is, 2011. Culture Communication Criticism. Jalasutra: Yogyakarta.
[2] Ayun, Pq, 2015. The Phenomenon of Adolescents Using Social Media in Forming Identities. Channel, 1-16.
[3] ROSYIDI., 2012. Personality Psychology Psychological Paradigm. Jaudar Press.
[4] Damanik, At, & Khodra, Ml, 2015. Predict the Personality of Big 5 Twitter Users with Support Vector Regression. Cybermatika, vol. 3 - no. 1.
[5] Negara, Abp, Muhardi, H., & Putri, Im, 2020. Analysis of Airline Sentiment Using the Naive Bayes Method and Selection of Information Gain Features. Journal of Information Technology and Computer Science, vol 7.
[6] Sarwani, Mz, & Mahmudy, Wf, 2016 Campus Sentiment Analysis E-Complaint Using Probabilistic Neural Network Algorithm. Cursors, 135 – 140.
[7] Sarwani, Mz, & Mahmudy, Wf, 2015. Twitter Analysis to Determine Someone's Character Using the Naïve Bayes Classifier Algorithm, SESINDO, 291-296.
[8] Shofa, Yasin & Rahmawati., 2015. Data Classification of Birth Weight Using Probabilistic Neural Network and Logistic Regression (Case Study at Sultan Agung Islamic Hospital Semarang, 2014).
[9] Sun, C., Hu Yi., & Shi, P., 2020. Probabilistic Neural Network Based Seabed Sediment Recognition Method For Side-Scan Sonar Imagery. Sedimentary Geology, vol 410.
[10] Alweshah, M., Rababa, L., Ryalat, Mh, Momani, A. A ,. & Ababneh, Mf, 2020. African Buffalo Algorithm: Training The Probabilistic Neural Network To Solve Classification Problem. Journal Of King Saudi University Computer And Information Sciences.
[11] FAN, H., PEI, J., & ZHAO, Y., 2020. An Optimized Probabilistic Neural Network With Unit Hyperspherical Crown Mapping And Adaptive Kernel Coverage. Neurocomputing, Vol 373.
[12] Lhaksamana, Km, Nhita, F., & Anggraini, D., 2017. Personality Classification Based on Facebook Status Using the Backpropagation Method. e-Proceeding of Engineering, 5174.
[13] Saputri, Ee, Wahono, Rs, & Suhartono, V., 2015. Selection of Smoothing Parameters on Probabilistic Neural Network by Using Particle Swarm Optimization for Text Detection on Images. Journal of Intelligent Systems.
[14] Sarwani, Mz, & Sani, Da, Fakhrini, FC, 2019. Personality Classification through Social Media Using Probabilistic Neural Network Algorithms, IJAIR, 9-15.
[15] Sarwani, Mz, & Sani, Da, 2020. Knowing Personality Traits on Facebook Status Using the Naïve Bayes Classifier, IJAIR, 22-28

DOI : 10.25139/inform.v6i1.3307