(IJID) International Journal on Informatics for Development, *e-ISSN: 2549-7448*
Vol. 7, No. 2, 2018, Pp. 59-63

Analysis of Personality Characteristic Using the Naïve Bayess Classifier Algorithm (Case Study Official Twitter of Basuki Tjahaja Purnama's and Anies Baswedan)

Ireicca Agustiorini Harsehanto
Student of Informatics Department, Faculty of Science and Technology, Universitas Islam Negeri Sunan Kalijaga
Yogyakarta, Indonesia, ireiccaah@gmail.com

M. Didik R. Wahyudi
Informatics Department, Faculty of Science and Technology, Universitas Islam Negeri Sunan Kalijaga
Yogyakarta, Indonesia, m.didik@uin-suka.ac.id

Abstract— This research uses data from social media Twitter based on the results of tweets from user_timeline @basuki_btp and @aniesbaswedan. This study uses 2100 tweet data. Data that has been collected is then pre-processed first and labeled manually. The next process is classification using the Naïve Bayes Classifier Algorithm using the Big Five Personality Theory. Based on the test results using 500 tweet data as training data and 1600 tweet data as testing data. The classification results obtained by using the Naïve Bayes Classifier Method and grouped in the "Big Five" personality groups: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism on tweet data in Indonesian.

Keywords-- Naïve Bayes Classifier (NBC); Social Media; Twitter; Personality Analysis; Big Five Personality

I. INTRODUCTION

One way to find out one's personality is to do a psychological test. Psychological tests conducted today are mostly through written tests or interview tests [1]. At present, there is a lot of research on a person's personality that is done using social media, because social media is a very important need for society, especially modern society today who always share their daily activities on social media [2].

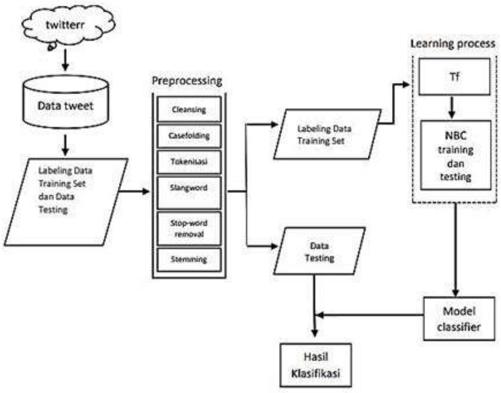
Social media is a medium to socialize with each other and be done online which allows humans to interact with each other without being limited by time and space. Social media at this time is very easy for users in various ways [3]. Not only to give important news, but social media can be used to share a short message, send messages, comment, make friends, send photos, space to exchange opinions and so forth [4]. It has been proven that in just a few years twitter has become one of the most popular social media among the community to date [5].

Twitter is also often used as a medium to publish daily activities through a short message consisting of a maximum of 280 characters (called tweets). Your own Tweet can consist of text and photo messages. Through this tweet, Twitter users can interact more closely with other users by sending about what they are thinking, what is being done, about the events that have just happened, about the latest news and other things so that many twitter users unconsciously provide information about his personality through tweets they made with natural language [2].

This research uses Twitter as a medium to analyze one's personality [2]. In the process of analyzing someone's personality through Twitter, it takes the right methodology to get accurate results. Tweet is a collection of words that are not standard so special treatment is needed to get data that can be processed. Therefore, in the processing of data, a pre-processing process is required, which can then be classified [6]. In this study, the method used is the Naïve Bayes Classifier and is grouped in the "Big Five" personality group [7]. The method was chosen because it is simple and provides convenience in the data processing process and provides an accurate level of accuracy [8].

II. RESEARCH METHODS

The research method used is an experimental method or applied research, namely by applying the Naïve Bayes Classifier method to analyze and classify a person's personality character in the "Big Five" personality group based on the original tweet on its user_timeline [9]. This study uses several Text Mining methods and Crawling data techniques for extracting data information from the source page [3]. Figure 1 describes the flow of the research process in general which consists of several stages, namely: data collection, data selection, data preprocessing and NBC processes and classification results [10].



Gambar 1. Alur Proses Penelitian

Figure 1 Research flows

1. Data Sources

Data source information is obtained from twitter.com. Collection data is done in a way crawling data with make use of the Twitter API [3]. Data needed in this research is tweet data native Indonesian language taken from Twitter user_timeline officially Basuki Tjahaja Purnama (@basuki_btp) and Anies Baswedan (@aniesbaswedan). The total amount of tweet data collected is as much 2100 data. Twitter data is then stored in the MySQL database.

2. Data Selection and Labeling

- Data Selection

Data tweets that have been collected cannot all be used for the process further research. This is because some tweets do not speak Indonesia, as well as several repeated tweets. Therefore, it is done selecting tweet data to determine the data to be used in the next stage. This selection process is only carried out on the data that will be used in the learning process. After going through the selection process, 2100 data were taken will be used in a supervised learning process [11].

- Data Labelling

Data labeling is not done on all data, but only done on the data that will be used in the training phase. Data that will be labeled manually this is 500 tweet data. This labeling is done manually and selected random, then each tweet will be grouped into five personality classes namely; Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A) and Neuroticism (N) [7].

3. Preprocessing

Tweet data that has been through a selection process is then carried out by the pre-processing process data. The steps in the pre-processing process are explained as follows:

- **Cleansing**

Data tweets usually contain html tags, url, hashtag (#), username, punctuation or other non-letter characters. At this stage, cleaning is carried out these components.

- **Case Folding**

At this stage, all letters in the tweet data are changed to lowercase. Existence capital letters will affect *tf* (Term Frequency) calculations. Therefore, case folding is done to generalize letters to lowercase.

- **Tokenizing**

Tokenizing is the process of dividing text, sentences or paragraphs into tokens or certain parts. Usually the separation of these tokens refers to presence of spaces or punctuation.

- **Slang-Words**

Sentences on twitter usually contain non-standard words that are not appropriate with standard Indonesian spelling. Therefore, at this stage it is carried out Word conversion, which is not standard into the standard EYD form (Enhanced Spelling).

- **Stop-Words Removal**

Stop Word is a common word that often appears normally ignored because it is considered to have no meaning. Examples of categorized words as stop word is a person pronoun and conjunction, such as ('I am', 'you', 'And', 'from', 'indeed', 'who', 'on', 'to' and others. This process is done with utilize the library stop word Sastrawi.

- **Stemming**

The process of converting words into documents into basic word forms uses Sastrawi stemming algorithm. Example: the word 'play' is changed to 'play', said ‘Explained 'changed to' clear 'and so on.

4. Personality Character Analysis Process

1) Word weighting with Term Frequency (TF)

TF (Term Frequency) is the frequency of the appearance of a term in a document concerned. The greater the number of occurrences of a term (high TF) in document, the greater the weight or the value of conformity the greater it is. Each word in the tweet table is then calculated frequency its appearance, then, calculates the value of the opportunities in each class personality with Formula (1):

$$P(i) = \frac{N_i}{N} \qquad (1)$$

Information:

P(i) = chance of sentiment i

$\boldsymbol{N_i}$ = number of document entered in sentiment i

N = total number of document

2) Naïve Bayes Classifier algorithm

The Naïve Bayes Classifier (NBC) algorithm is one of the classification algorithms. This algorithm uses a machine learning method that utilizes probability and statistics put forward by British scientist Thomas Bayes. The main characteristic of the algorithm Naive Bayes Classifier is a very strong (naïve) assumption of independence from each condition / event [12]. The NBC algorithm has better accuracy than with other classifier models Bayes theorem is the basic rule of the following Naïve Bayes Classifier Bayes equation theorem given in Formula (2):

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad (2)$$

Where *X* is the tuple data of the test results from a predetermined data set enter a particular class. H is a hypothesis that will determine *X* enter into class *C*. *P* (*H* | *X*) is an opportunity or probability of X which is tuple data or evidence obtained when observation enters class *C*, with other words looking for probability *X* is owned by class C. P (H | X) is a probability posterior, *H* is conditioned on *X*. Instead, *P* (*H*) is the prior probability, or previous probability. Then *P* (*X* | *H*) is the posterior probability where *X* conditioned on *H*. While *P* (*X*) is the previous probability of *X*. As for this research will implement Bayes rules that are stated as follows:

$$P(Cj|X) = \frac{P(X|Cj)P(Cj)}{P(X)}$$

Where *Cj* is the category of text to be classified, and *P* (*Cj*) is prior probability of the text category *Cj*. Whereas d is a text document represented as a set of words (*W1*, *W2*, ... *Wn*), where *W1* is the word first, *W2* is the second word and so on. During the process of classifying text documents, the Bayes approach will choose the category that has the highest probability (*C MAX*), namely *C MAX* given in Formula (3):

$$C_{MAP} = argmax \frac{P(X|Cj)P(Cj)}{P(X)} \qquad (3)$$

The value *P* (*X*) can be ignored because the value is constant for all *Cj*, so the above equation can be written:

$$C_{MAX} = argmax\, P(X|Cj)P(Cj)$$

Probability *P* (*Cj*) can be estimated by calculating the number of training documents on each category of *Cj*. In other hand, to calculate the distribution of *P* (*X* | Cj) it will be difficult because the number of terms becomes very large. This is because the number of terms is the same as the sum of all word position combinations multiplied by the number of categories that will classified. With the Naïve Bayes approach, that assumes that every, the words in each category are independent of one another, and then the calculation can be simplified and can be written as follows:

$$P(X|Cj)\prod_{i=1}^{n} P(W_i|C_j)$$

By using the *C MAP* equation, the equation can be written as:

$$C_{MAP=argmax\ (Cj)}\prod_{i=1}^{n} P(W_i|C_j)$$

The *P* (*Cj*) and *P* (*Wi* | *Cj*) values are calculated during the training process where the equation is as follows:

$$P(Cj)=\frac{docs\ j}{contoh}$$

$$P(W_i|C_j)=\frac{1+n\ i}{|C|+n\ (kosakata)}$$

Information:
P (*Wi* | *Cj)* = Probability of the word *Wi* in the *Cj* category
| docs *j* | = Number of documents in category *j*
| example | = The number of all sample documents used in training process
ni = Frequency of occurrence of the word *Wi* in the category *Cj*
| *C* | = Number of all words in the *Cj* category
n (vocabulary) = Number of unique words in all training data

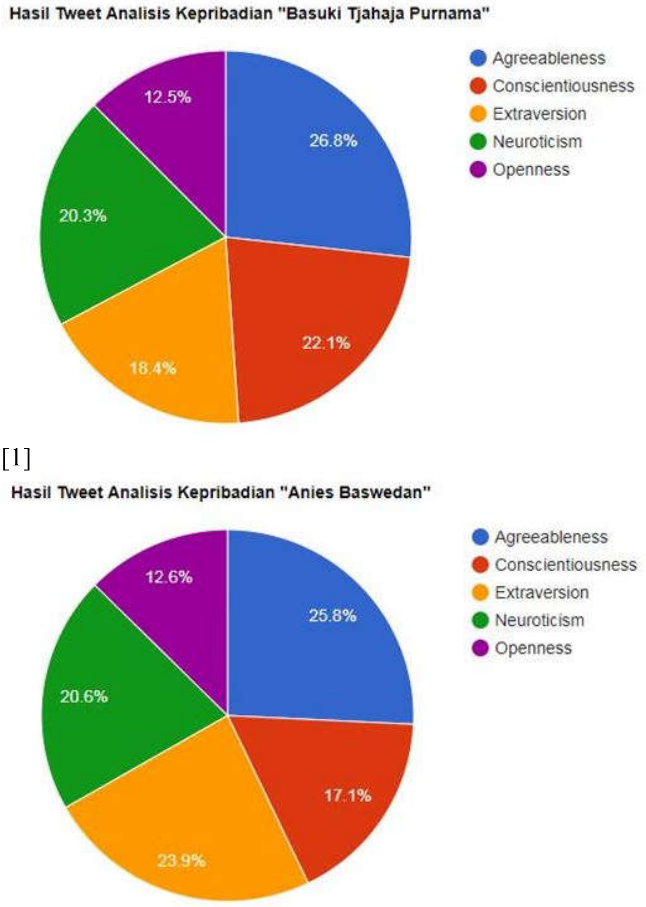
III. EMPIRICAL RESULTS

In this process, the implementation phase is carried out using the Naïve Bayes Classifier method and is classified into groups of Big Five Personality. Data will be used in the learning process as many as 2100 tweet data obtained from user_timeline twitter @between_btp and @aniesbaswedan. Each has a number the same data as presented in Table 1.

Table 1 Number of Tweets

No	User_Name	Jumlah tweet
1	@basuki_btp	1050
2	@aniesbaswedan	1050
Total		2100

In the learning process, data is divided into two, namely 500 data created as training data and the remaining 1600 is used as data testing for accuracy testing. Results of classification use the Naïve Bayes Classifier method and are grouped in & quot; Big Five & quot;personality group: Openness, Conscientiousness, Extraversion, Agreeableness,Neuroticism in tweet data in Indonesian. Results obtained from classification can be seen and explained in a circle diagram. The diagram image can be seen in Fig. 2.



[1]

Figure 2. Result of classification characteristic personality diagram

From Fig. 2, it can be concluded that the results show the most trait big or dominant at the same time is in Agreeableness and the smallest trait is at Openness, the most distinguishing in the second trait is Conscientiousness for Basuki and Extraversion for Anies.

IV. CONCLUSION

Based on research that has been done using the Naïve Bayes Classifier method and grouped in & quot; Big Five & quot; Personality Group (Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism) in the inner tweet data Indonesian language is concluded as follows:

1) Basuki Tjahaja Purnama gets the biggest or dominant trait value on Agreeableness is 26.8%, then Conscientiousness is 22.1%, Neuroticism is 20.3%, 18.4% extraversion and 12.5% for openness.
2) Anies Baswedan obtained the biggest or dominant trait value on Agreeableness is 25.8%, then Extraversion 23.9%, Neuroticism 20.6%, 17.1% Conscientiousness and 12.6% for Openness.

REFERENCES

[1] Claudy, Y.I., Perdana, R.S., & Fauzi, M.A , "Klasifikasi Dokumen Twitter untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritme K-Nearest Neighbor (KNN)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 2(8), pp. 2761-2765, 2018.

[2] Sarwani, M.Z., & Mahmudy, W.F, "Analisis Twitter untuk Mengetahui Karakter," *Seminar Nasional Sistem Informasi Indonesia,* pp. 2-3, 2015.

[3] Valkanas, G., Saravanou, A & Gunopulos, "A Faceted Crawler for the Twitter Service," 2014. [Online]. Available: https://doi.org/10.1007/978-3-319-11746-1_13.

[4] Kwak, H., Lee, C., Park, H., & Moon, S, "What is Twitter, a social network or a news media?, 591," 2010. [Online]. Available: https://doi.org/10.1145/1772751.

[5] Zarrella, D., "The Social Media Marketing Book," 2014. [Online]. Available: https://doi.org/10.1007/s13398-014-0173-7.2.

[6] Y. E. Zohar, Introduction to Text Mining, University of Illinois, 2002.

[7] Barrick, M.R., & Mount, M.K, "the Big Five Personality Dimensions and Job Performance: a Meta- Analysis," *Personnel Psychology,* vol. 44(1), pp. 1-26, 1991.

[8] Feist, J., Feist, G, Theories of Personality, Jakarta: Salemba Humanika, 2010.

[9] Pascasarjana, P., & Udayana, U, Text Mining dengan Metode Naive Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis, Bali: Universitas Udayana, 2011.

[10] Hidayatullah, A. F., Sarjadi, S, Analisis Sentimen dan Klasifikasi Kategori terhadap tokoh publik pada twitter, Yogyakarta: Universitas Gajah Mada, 2014.

[11] Buntoro, G.A. , "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter.," 2017.

[12] Natalius, S, Metode Naive Bayes Classifier dan Penggunaannya pada Klasifikasi Dokumen, 2010.