OPEN ACCESS
J DATA SCI APPL, VOL. 1, NO. 1, PP.057-066, JULY 2020

E-ISSN 2614-7408
DOI: 10.34818/JDSA.2020.3.61

JOURNAL OF DATA SCIENCE AND ITS APPLICATIONS

Classification of Personality based on Beauty Product Reviews Using the TF-IDF and Naïve Bayes (Case Study : Female Daily)

Novia Russelia Wassi[1], Adiwijaya[2], Mahendra Dwifebri Purbolaksono[3]

School of Computing, Telkom University
Bandung, Indonesia

* adiwijaya@telkomuniversity.ac.id

Received on 18-02-2021, revised on 10-06-2021, accepted on 17-06-2021

Abstract

A person's personality is an important parameter to determine the character of each person and as an assessment in various ways. Currently personality can not only be known from psychological tests, but also can be known in various ways. One way is through reviews presented in electronic media. Therefore, in this study we can know the process of Classification of Personality based on Beauty Product Reviews and know the results of the classification. In this study, a person's personality was classified into three "Big Five" personality groups, namely: Openness, Conscientiousness, and Extraversion using the Naïve Bayes method and TF-IDF as Feature Extraction. The results of the classification that have been done get 81% accuracy with preprocessing scenarios using Stemming and Stopword, TF-IDF unigram, and BernoulliNB classifier type.

Keywords: big five personality, female daily, feature extraction, naïve bayes, TF-IDF

I. INTRODUCTION

Personality is a behavior that reflects how the character of each person can be used as a benchmark for assessment in any case, for example an assessment in terms of school, work and so on. There are many classification methods in personality, one of the most frequently used methods is the Big Five Personality method, namely: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism.[1]. There are many methods to find out the personality of everyone, through psychological tests and interviews. But to do a psychological test is not easy, everyone must prepare a lot of money just to find out each other's personality. Therefore, personality analysis and classification can be carried out only with reviews written by individuals on the Female Daily forum.

Female Daily is a forum that contains reviews from visitors for various kinds of beauty products. In this forum, you can find various kinds of individuals who write reviews according to how they feel after using these beauty products. With the review data that has been written by these visitors, it can be used to determine the personality of each visitor on the Female Daily forum.

In this final project research, will perform personality classifications based on visitor reviews using the Naïve Bayes algorithm. Naïve Bayes itself is a classification technique that has progressed rapidly and has become a core technique in classification[2]. This algorithm focuses on distribution assumptions that are made based on many occurrences of words[2]. Naïve Bayes Classifier also simple, fast in making decisions, easy to apply and does not require large amounts of data[3], therefore in this study chose to use the Naïve Bayes method rather than other methods. In this study also using word weighting with the TF-IDF method. TF-IDF is an algorithm used to calculate the weight of each word and the number of times the word appears in a text[4]. So it can be said that TF-IDF is the most effective method for calculating word weighting[5].

This study aims to classify personality using the Naïve Bayes method and TF-IDF as Feature Extraction. Personality classes are divided into three parameters of the Big Five Personality, namely: Openness, Conscientiousness, and Extraversion. The results obtained from each classification are the level of these parameters, for example:

74,3545: 1,1,0

In the data above, the data with id 74.3545 has Openness and Conscientiousness parameters in beauty product reviews and does not have Extraversion parameters. Prior to classification, a preprocessing process is carried out including: data cleaning, case folding, remove punctuation, remove numbers, data normalization, remove meaningless, stemming and filtering, weighting TF-IDF as feature extraction.

This study contains 4 chapters after the Introduction, chapter 2 Literature Review, Chapter 3 Research Method, Chapter 4 Results and Discussion and Chapter 5 Conclusion.

II. LITERATURE REVIEW

A. Big Five Personality

Many methods can be used to determine the personality of one individual and another, one of which is psychological test. In psychological tests, there are also many theories that can be used to measure the individual's personality. One of the most widely used theories is the Big Five Personality theory[6]. What is meant by the Big Five Personalities, namely: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism. The following is an explanation of each group of Big Five Personalities:

1) Openness: This trait is an individual's personality that is open to new experiences. Individuals who have this trait tend to be more creative and imaginative.

2) Conscientiousness: This trait is a careful individual personality. Individuals who have this trait tend to be disciplined, responsible, diligent and reliable.

3) Extraversion: This trait is an individual personality who is easily comfortable with the conditions around him. In other words, this trait is an adaptable individual personality.

4) Agreeableness: This trait is an individual personality who easily agrees with the opinions of others. Individuals who have this trait tend to be more trustworthy, cooperative and soft-hearted.

5) Neuroticism: This trait is an individual personality who is not easily depressed in dealing with problems. Individuals who have this trait tend to be calm in dealing with problems, confident and have a firm stand.

B. Female Daily

Female Daily is a forum that contains reviews from visitors for various kinds of beauty products. In this forum, you can find various kinds of individuals who write reviews according to how they feel after using these beauty products. With the review data that has been written by these visitors, it can be used to determine the personality of each visitor on the Female Daily forum.

C. Naïve Bayes Classifier

The Naïve Bayes Classifier is a simple classification method based on the Bayes theorem using probability[7]. Naïve Bayes Classifier assumes that the variable is a variable that is independently assigned by the class[7]. Naïve Bayes Classifier has high accuracy when applied to databases with data[8]. The advantage of the Bayes Classifier is that this method is simple, fast in determining decisions, easy to apply, and does not require large amounts of data[8]. The stages of the Bayes Classifier process are counting the number of classes. After knowing how many classes, count the number of cases from each class. Then, multiply all the variables in the class and compare the results between each class. Mathematically, the formula for the Naïve Bayes Classifier equation is as follows[9]:

$$P(C|X) = \frac{P(x|c)P(c)}{P(x)}$$

Information:

x	: Data with unknown class.
c	: Data testing which is the classification result classes.
P(C|X)	: Probability of hypothesis c based on condition x.
P(x|c)	: Probability based on condition x in hypothesis c.
P(c)	: Probability hypothesis c.
P(x)	: Probability x.

D. Feature Extraction

The feature extraction process is needed in the classification. The purpose of feature extraction itself is to weight words by calculating the weight of each word and the number of times the word appears in a text.[10]. The feature extraction algorithm used in this study is the TF-IDF algorithm. TF-IDF is one of the most widely used algorithms. Term Frequency itself means the number of times a word is repeated in a text. Meanwhile, Inverse Document Frequency is an algorithm that is used to calculate the probability of searching for a word in the text. The equation for the TF-IDF method is as follows:

$$TF * IDF(d,t) = TF(d,t) * log\frac{N}{df(t)}$$

Information:

TF * IDF(d, t)	: Weighted TF-IDF.
TF(d,t)	: Frequency of appearance of term t in the document d.
N	: The sum of all documents.
df(t)	: Number of documents containing the term t.

III. RESEARCH METHOD

In this research, the process carried out is preprocessing including: data cleaning, case folding, remove punctuation, remove numbers, word normalization, remove meaningless, stemming and filtering, weighting TF-IDF as feature extraction, classification with Naïve Bayes and evaluation. The following is an overview of the system scheme that will be built into the top five personality classifications based on beauty product reviews:

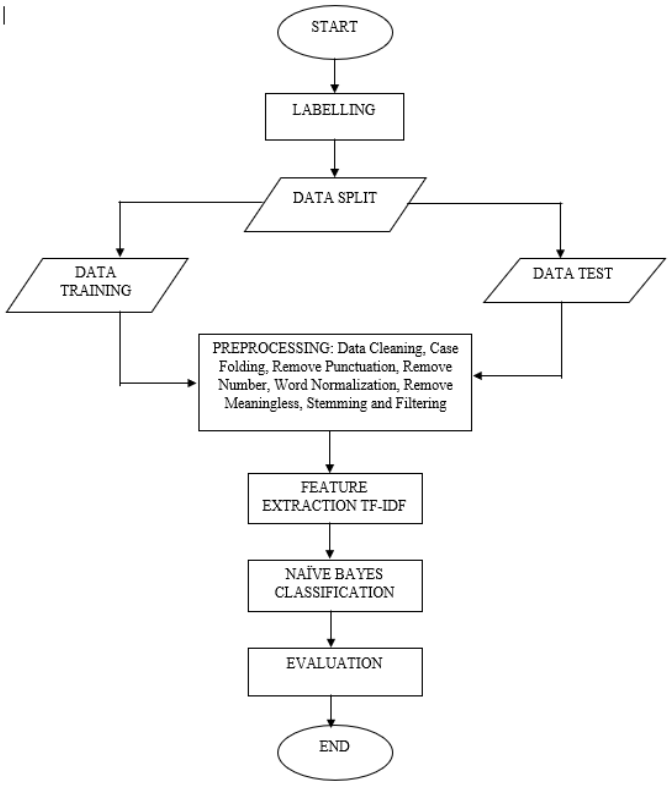

Fig. 1. System Schematic Overview

A. Dataset

The dataset that will be used in this final project research is data from the Female Daily forum. The data used is 500 data and focuses on data with the review_text attribute or visitor reviews with the skincare category. Included in this skincare category include Facial Wash, Toner, Wash-Off, Serum & Essence, Face, Sun Protection and Scrub & Exfo.

B. Labelling

At this stage, 3-dimensional labeling of the Big Five Personality is carried out, namely Openness (O), Conscientiousness (C), and Extraversion (E) on the available dataset from female daily forums. The review_text attribute is labeled according to the properties of the 3 dimensions that have been determined[11]. Here's an example of labelling:

TABLE I
LABELLING EXAMPLE

product_category	review_text	O	C	E
Facial Wash	"enak banget produk ini pas di pake ga bikin muka kering, dulu pernah pake facial wash yg klo abis cuci muka itu bner2 kering banget smpe suka ada putih2 di muka, pas pake hada labo enak hydrating banget buat kulit.."	0	1	1
Toner	"setelah pencarian sekian lama, akhirnya ketemu juga ama toner yang ga bikin purging dan teksturnya cair. Sebelumnya aku pakai seminggu sekali, soalnya takut kalaunya pemakaian berlebih bisa sampai dehidrasi iritasi bahkan sampai breakout lagi (soalnya pernah sekali), ini bener-bener magic sih, pori-pori jadi lebih kalem, bekas-bekas jerawat jadi makin berkurang, muka lebih bersih daripada sebelumnya pakai bha ini, seneng banget!!! bakal repurchase terus mungkin bakal mau coba yang gel, katanya lebih hemat yang gel sih daripada yang liquid.."	1	1	1

The labeling factor follows the reference of several previous studies, Openness personalities tend to prefer to try new things that have not been tried before and consider it as a challenge[11]. The Conscientiousness personality always pays attention to details, individuals with the Conscientiousness personality will see things and explain everything in detail[11]. While the Extraversion personality is a talkative personality, in the sense that it likes to tell everything that has happened, an individual with this personality is also someone who is expressive[11]. The form of labeling is to give the notation 0 and 1, with 0 which means it does not have the personality in question. Whereas 1 means having the personality in question.

TABLE II
PERCENTAGE OF CLASSES IN EACH COMBINATION

Openness	*Conscientiousness*	*Extraversion*	Presentase
-	-	-	0,4%
-	-	Y	6,8%
-	Y	-	22,6%
-	Y	Y	37,2%
Y	-	-	5%
Y	-	Y	7,4%
Y	Y	-	3,2%
Y	Y	Y	17,4%

In the class percentage of each combination, it was found that the highest percentage was reviews with a combination of personality Conscientiousness and Extraversion with 37.2%. Meanwhile, the lowest percentage were reviews that did not contain any of the 3 dimensions, with 0.4%.

C. *Data Split*

In the data split process, the sharing process is carried out between the Training Data and Test Data. The comparison used is 80:20, where 400 of training data and 100 of test data. This process also determines "random_state = 42". The value from the random state will provide a different set of training data and test data, and the value 42 is the best set because it produces the best accuracy. The random state is defined to make the value consistent the next time the system restarts.

D. *Preprocessing*

In this process, a preprocessing stage is carried out which aims to obtain data that is ready to be processed in the classification process. The data obtained will be simpler than the original dataset. The process carried out is:

1. Data Cleaning, this process removes symbols on the document, with the aim of cleaning the document.
2. Case Folding, this process changes the text in the document to lowercase or lowercase.
3. Remove Punctuation, this process removes any punctuation marks on the document.
4. Remove Number, this process removes the numbers in the document.
5. Data Normalization, this process converts non-standard words into standardized words, by creating your own data dictionary.
6. Remove Meaningless, this process removes unnecessary words in the document, by creating your own data dictionary. Examples of unimportant words are 'otw', 'oh', 'ohiya', and so on.
7. Filtering, this process removes unnecessary words, namely the words included in the stopword. The words included in the stopword are "di", "yang", "dan", "dari", etc[12]

E. *TF-IDF*

In TF-IDF, there is one parameter, namely N-gram. The N-gram on the TF-IDF is responsible for classifying the features of the document. There is a Unigram and Bigram division. Unigram groups the features word by one word and Bigram groups the features word by two words[13].

F. *Naïve Bayes*

This study uses three types of classifiers in the Naïve Bayes Classifier, namely BernoulliNB, GaussianNB, and MultinomialNB. BernoulliNB is a good classifier for features that have binary values[14]. MultinomialNB is a classifier commonly used in multiclass data by storing many words that often appear in documents.[14][15]. Meanwhile, GaussianNB is a classifier that is usually used when features have a continuous value[16].

G. *Evaluation*

The last process is performance measurement. Performance measurement is carried out to determine the accuracy of the algorithm used, namely, Naive Bayes. There are several parameters that can be used, namely Recall, Precision, and F-Measure[17]. To make it easier to calculate performance with existing parameters, you can use the Confusion Matrix. The following is a table of the confusion matrix[17]:

TABLE III
CONFUSION MATRIX

		Actual Class	
		class = no	class = yes
Predicted Class	class = no	TN	FP
	class = yes	FN	TP

Information:

TP : True Positive, is a class that predicted yes, but in reality it is also yes.
FN : False Negative, is a class that predicts no, but in fact it is also no.
FP : False Positive, is a class that predicted no, but in fact yes.
TN : True Negative, is a class that predicted yes, but in fact it is no.

1. Recall
   This parameter counts from all positive classes, how many we predict is correct[17].

$$\boldsymbol{Recall = \frac{TP}{TP + FN}}$$

2. Precision
   This parameter counts out of all the positive classes that we predict to be correct, how many are really positive[17].

$$\boldsymbol{Precision = \frac{TP}{TP + FN}}$$

3. F-Measure
   This parameter computes recall and precision simultaneously, to make the two parameters comparable[17].

$$\boldsymbol{F - Measure = \frac{2 * recall * precision}{recall + precision}}$$

IV. RESULTS AND DISCUSSION

A. Result

In this study, use 500 data with 400 data for data training and 100 data for data testing. It also use 3 combinations of test scenarios were carried out to obtain the model that has the best performance. The first scenario is to compare the preprocessing process, namely by doing the Stemming and Stopword Removal processes, without doing the Stemming process, and without doing the Stopword Removal process. The second scenario is the classifier used in the classification process, namely BernoulliNB, GaussianNB and MultinomialNB. While the third scenario is to compare unigram and bigram words on TF-IDF.

B. Analysis Results

1. The Results of the Preprocessing Analysis
   The test results show that the Stopword Removal process can cause the accuracy to decrease. The Stopword Removal process is removing words that are included in the Indonesian NLTK dictionary. For example, a sentence that should be “tidak membuat” if Stopword Removal is performed will become “membuat”, and that will make the sentence have a different meaning. In addition to the Stopword Removal Process, a Stemming process is also carried out which converts words that have affixes into basic word forms. For example, the word “menghidrasi” becomes “hidrasi”. The Stemming process is carried out, the accuracy will be higher than the Stemming process is not carried out, this is because if a word is not carried out by the Stemming process, then the affixed word and the root word will have a different meaning, even though in reality the meaning is the same.

TABLE IV
THE RESULTS OF THE PREPROCESSING COMPARISON ACCURACY

Stopword	*Stemming*	*Macro Precision*	*Macro Recall*	*Macro* F1 *Score*	*Max Macro* F1 *Score* (%)
Y	Y	0,79	0,84	0,81	
Y	-	0,71	0,75	0,73	81%
-	Y	0,78	0,76	0,77	

It can be seen from the table that the accuracy results without Stopword Removal are higher than Stopword Removal, which is 77%. Meanwhile. The results of the accuracy by not doing Stemming decreased by 73%. However, if we do a combination of Stopword Removal and Stemming, the resulting accuracy will be maximal at 84%, because the preprocessing process is more complete.

2. Classifier Comparison Analysis Results

The classifiers used are BernoulliNB, GaussianNB and MultinomialNB. The results obtained from the three classifiers are that BernoulliNB has the highest accuracy compared to the other two classifiers.

TABLE V
CLASSIFIER COMPARISON ACCURACY RESULTS

Classifier	*Macro Precision*	*Macro Recall*	*Macro* F1 *Score*	*Max Macro* F1 *Score* (%)
BernoulliNB	0,79	0,84	0,81	
GaussianNB	0,71	0,74	0,71	81%
MultinomialNB	0,69	0,69	0,64	

BernoulliNB can have the higher accuracy of 81% compared to the other two classifiers because BernoulliNB focuses on features that have binary values. As we know, the value of each label only uses the notation 0 and 1, so the most suitable classifier is BernoulliNB.

3. The Results of the Comparative Analysis of the Word Unigram and Bigram.

The N-gram on the TF-IDF is responsible for classifying the features of the document. There is a Unigram and Bigram division. Unigram group word by one word features and Bigram group word by two words features and this will cause many word features that are not present in the training data.

TABLE VI
THE RESULTS OF THE ACCURACY OF THE COMPARISON OF UNIGRAM AND BIGRAM WORDS

N-gram	*Macro Precision*	*Macro Recall*	*Macro* F1 *Score*	*Max Macro Score* (%)
Unigram	0,79	0,84	0,81	81%
Bigram	0,65	0,74	0,68	

The results obtained from this research, Unigram has a higher accuracy than Bigram. Unigram has the highest accuracy of 81% and Bigram has the highest accuracy of 68%. Bigram has lower accuracy than Unigram because the word features are more varied. For example, in the sentence “enak banget

produk ini pas di pake ga bikin muka kering" if you do the Bigram process it will be "enak banget", "banget produk", "produk ini", "ini pas", "pas di", "di pake", "pake ga", "ga bikin", "bikin muka" and "muka kering". Many features are irrelevant in this example, one of which is "banget produk" and this will decrease the accuracy.

V. Conclusion

The conclusion obtained from the results of this study is that how to classify the big five personalities based on beauty product reviews is by doing Labeling, Split Data using the Train / Test Split Method, Preprocessing (Data Cleaning, Case Folding, Remove Punctuation, Remove Number, Word Normalization, Remove Meaningless, Stemming and Filtering), Feature Extraction using TF-IDF, Classification using the Naïve Bayes method and Evaluation. Based on the results of the tests that have been done, it is found that preprocessing without doing Stopword removal will make the accuracy higher, which is 77% than not doing. Stemming will make the accuracy decrease to 73%. However, if the two preprocessing processes are carried out, the accuracy is more than 81%. Using the BernoulliNB classifier will make the resulting accuracy better than other classifiers which is 81%, because BernoulliNB focuses on features with binary values. In addition, using Bigram word will make the accuracy decrease due to the division of word features by two words, so that it will make word features to be absent in the training data and words become irrelevant.

Suggestions for further research to be able to improve accuracy are expected to carry out a more complete preprocessing process and also carry out a feature selection process so that only the selected features are used and of course the results will be maximized.

ACKNOWLEDGMENT

I thank the one and only God who has given me the wisdom to be able to complete this research. I also thank Mr. Adiwijaya and Mr. Mahendra as my mentors. Thanks also to all my friends who always want to help when there are difficulties.

REFERENCES

[1] "Teori Kepribadian Model Lima Besar (Big Five Personality) – IPQI." [Online] Available at: https://ipqi.org/teori-kepribadian-model-lima-besar-big-five-personality/ [Accessed 3 April 2020].
[2] Lewis, D. D. (1998). N a i v e (B a y e s) at Forty : T h e I n d e p e n d e n c e A s s u m p t i o n in I n f o r m a t i o n R e t r i e v a l. *Ecml*, *x*, 16.
[3] Zulfikar, W. B., Irfan, M., Alam, C. N., & Indra, M. (2017). The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter. *2017 5th International Conference on Cyber and IT Service Management, CITSM 2017*. https://doi.org/10.1109/CITSM.2017.8089231
[4] Dadgar, S. M. H., Araghi, M. S., & Farahani, M. M. (2016). A novel text mining approach based on TF-IDF and support vector machine for news classification. *Proceedings of 2nd IEEE International Conference on Engineering and Technology, ICETECH 2016*, *March*, 112–116. https://doi.org/10.1109/ICETECH.2016.7569223
[5] Kuang Q, Xu X. (2010). Improvement and Application of TF · IDF Method Based on Text Classification. *2010 International Conference on Internet Technology and Application*.
[6] Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1
[7] Soria, D., Garibaldi, J. M., Ambrogi, F., Biganzoli, E. M., & Ellis, I. O. (2011). A "non-parametric" version of the naive Bayes classifier. *Knowledge-Based Systems*, *24*(6), 775–784. https://doi.org/10.1016/j.knosys.2011.02.014
[8] Zulfikar, W. B., Irfan, M., Alam, C. N., & Indra, M. (2017). The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter. *2017 5th International Conference on Cyber and IT Service Management, CITSM 2017*. https://doi.org/10.1109/CITSM.2017.8089231
[9] "Algoritma Naive Bayes – Informatikatologi." [Online] Available at: https://informatikalogi.com/algoritma-naive-bayes/ [Accessed 3 April 2020].
[10] Park, M. S., & Choi, J. Y. (2009). Theoretical analysis on feature extraction capability of class-augmented PCA. *Pattern Recognition*, *42*(11), 2353–2362. https://doi.org/10.1016/j.patcog.2009.04.011
[11] "The Big Five Personality Traits." [Online] Available at: https://www.verywellmind.com/the-big-five-personality-dimensions-2795422 [Accessed 13 December 2020].
[12] "Text Preprocessing – Informatikatologi." [Online] Available at: https://informatikalogi.com/text-preprocessing/ [Accessed 9 April 2020].
[13] Hamzah, A. (2010). Deteksi Bahasa Untuk Dokumen Teks. *Seminar Nasional Informatika*, *22*(semnasIF), 5–13.
[14] "Bernoulli Naïve Bayes." [Online] Available at: https://iq.opengenus.org/bernoulli-naive-bayes/ [Accessed 1 February 2021].
[15] Gerard Maas, Francois Garillot. 2019. Stream Processing with Apache Spark. USA: O'Reilly Media, Inc(page. 379).
[16] Aboul Ella, Ahmad Taher, Tarek Gaber, Roheet Bhatnagar, Mohamed F. Tolba. 2019. Advances in Intelligent Systems and Computing. Switzerland: Springer Nature Switzerland AG(page. 663).