Personality Types Classification for Indonesian Text in Partners Searching Website Using Naïve Bayes Methods

Ni Made Ari Lestari[1], I Ketut Gede Darma Putra[2] and AA Ketut Agung Cahyawan[3]

[1]Department of Information Technology, Udayana University
Bali, 80119, Indonesia

[2]Department of Information Technology, Udayana University
Bali, 80119, Indonesia

[3]Department of Information Technology, Udayana University
Bali, 80119, Indonesia

Abstract
The development of digital text information has been growing fast, but most of digital text is in unstructured form. Text mining analysis is needed in dealing with such unstructured text. One of the activities important in text mining is text classification or categorization. Text categorization itself currently has a variety of approaches such as probabilistic approaches, support vector machines, and artificial neural network or decision tree classification. Naive Bayes probabilistic method has several advantages of simplicity in computing. Naïve Bayes method is a good method in machine learning based on training data using conditional probability as the basic. This experiments use text mining with Naïve Bayes method to classify the personality type of user and use the type to find their couples based on the compatibility of their personality type.
Keywords: *text mining, classification, personality, naive bayes*

1. Introduction

Development of science and computer technology has given an enormous influence in Information technology's world, thereby encouraging the appearance of various types applications, such as desktop, web, or mobile. Among the three applications, web is the most rapidly progressing now, that's make internet has become a primary requirement. Percentage of internet users today is very large. Almost all people know and use the internet for daily needs. Starting from simple things such as communication, social networking to business. About 85% of the data available on the internet has an unstructured format, so it needs to be developed a system that is able to automatically categorize and classify the data is not structured [1]. Automatic text categorization is one of the solutions to the problem because they can significantly reduce the cost and time manual categorization. The abundance of information unstructured text has encouraged the appearance of a new discipline in text analysis, namely text mining that tries to find patterns of information that can be extracted from a text that is not structured. By that understanding the text mining term refers also to the text data mining (Hearst, 1997) or knowledge discovery from text databases (Friedman and Dagan, 1995). Text mining can provide a solution to the problem of processing, organizing, and analyzing the unstructured data in large numbers. According to Saraswati (2011), the current text mining has gained attention in many areas, such as security application, biomedical applications, software and applications, online media, marketing applications, and academic applications. [2]

Documents classification based on similarities features or content of the document. Classification is done by entering documents into categories predetermined. That classification method is called supervised learning. Generally, the method of classification divided into two, are supervised learning and unsupervised learning. First, supervised learning is a method of grouping documents, which class or category of documents predefined; whereas unsupervised learning is clustering documents automatically without define a category or class first. [3]

From numerical based approach group, Naïve Bayes has several advantages such as simple, fast and high accuracy. Naïve Bayes for classification or categorization of text using word attributes that appear in a document as a basis for classification. Some research showed that although the assumption independence between words in a document is not fully met, but performance in the NBC classification is relatively very good. Previous experiments results showed the accuracy of Naive Bayes is to reach 90%.[4]

Allport (1937) defined Personality as the dynamic organization within the individual of those psychophyscal systems that determine his unique adjustment to his environment. Temperament appears from our genetic endowment and influences or is influenced by the experience of each individual, and one of its outcomes is the adult personality [5]. There are many theories about personality. The most commonly known personality theory is the theory of the four temperaments from Hippocrates. Hippocrates divided the human temperaments into 4 big categories. Each category can be mixed and have a dominant trait in the

human body and form a mixed personality. It also has a match temperament between temperaments that can be used to determine a match between human beings who have different temperaments. [6]

2. Previous Research

Research related to text mining using Naïve Bayes method with several research objects as follows:

The study entitled Application of Naive Bayes for classification SMS Customer's Voice (Case Study PT. Pertamina UPMS V Surabaya). The raised issued is how to implement Naive Bayes in classifying SMS customers voice into categories determined by PT. Pertamina UPMS V Surabaya and classify SMS customers voice based on department which is determined by PT. Pertamina UPMS V Surabaya. In the Naive Bayes algorithm, SMS data voice subscribers in the past, will be entered for the training process that will result in probabilistic models. This research use 40 learning document and 40 classification document. And the result for accuration rate is 97,5%. [7]

The study entitled text Mining with Naïve Bayes Method Classifier and Support Vector Machines for Sentiment Analysis. Test performed to compare the use of Naive Bayes and SVM for Sentiment Analysis. Sentiment Analysis is a computational studies of the opinions of people, appraisal and emotion through entities, events and attributes owned (Biu, L. 2010). In this research is used the Indonesian and English documents. Each data has positive and negative values, each of which will be tested by the method of NBC and SVM. From the test results that the SVM can provide good results for the positive test data and NBC gave good results for the negative test data. [8]

The study entitled Text Document Keywords Extraction Using Naïve Bayes Method. Tests conducted to determine the influence the use of two features (TFxIDF and PD) and 4 features (TFxIDF, PD, PT, and PS) on the accuracy of the system generated keywords. The first test conducted on 10 documents at 20, 30, and 40 documents training with stopwords elimination. The second test performed on 10 documents of different tests at 20 and 30 training documents the elimination of stopwords. Then see the results by preccicion values, recall, and f-measure. The result is training documents provide the value of certain tendencies how should the value of the features of the keyword, with more and more features that use the word (which is the keyword) the value of the probability becomes greater keywords and words (which rather than keywords) decreases the probability values. [9]

The study entitled Spam Email Classification with Naïve Bayes Classifier Method use Java Programming. This study tested the validity of a document whether or not including spam. The accuracy of the test results obtained the error rate when categorizing spam use NBC. The biggest error rate is when the training data used reaches 40. That is because the difference the number of keywords in the second category is too much. So that lead to a greater level of error than others. [10]

3. Methodology

The overview diagram of this research is shown in Figure 1.

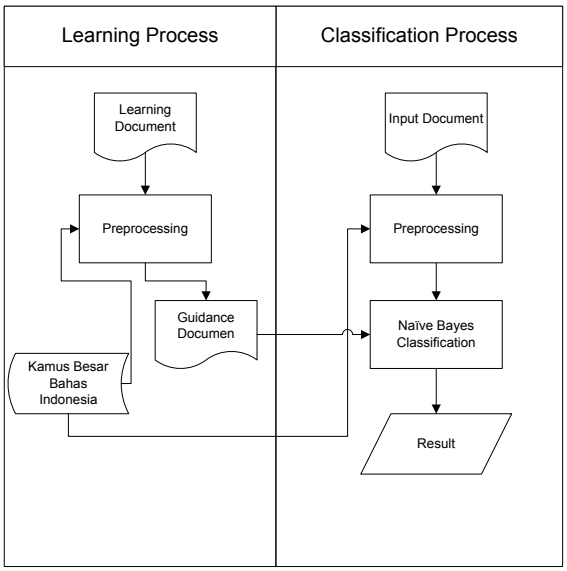

Figure 1 Research Overview Diagram

3.1 Preprocessing Text

Text preprocessing phase has been showed in figure 2.

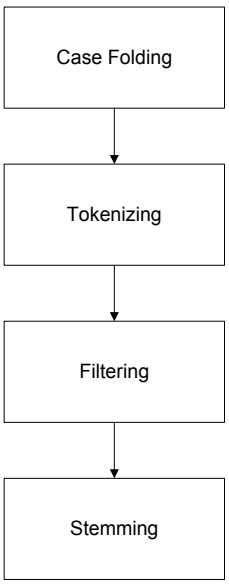

Figure 2 Text Processing Step

1. Case folding is the phase of changing uppercase to lowercase in the document then the elimination of

punctuation other than the "a" to "z" letter which is considered as the delimiter character.

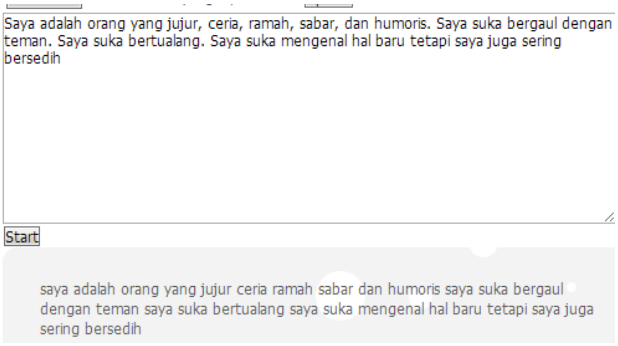

Figure 3 Case Folding Process

2. Tokenizing is the phase of splitting sentence to words. With the word's splitting first, the string that has been input will be simpler because showed in each words according to space which split it, so with that form, will make easier the changing process to be a word stem.

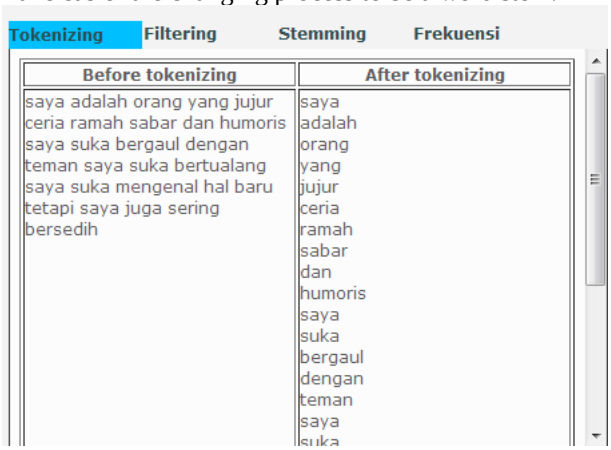

Figure 4 Tokenizing Process

3. Filtering is the phase of removal the words is not considered contain any meaning or thought there should be exist (Stopwords). Words in the stopwords list must be removed.

Figure 5 Filtering Process

4. Stemming is the phase of disposal affixes the words, either a prefix or a suffix. The flowchart for stemming process as seen in figure 6.

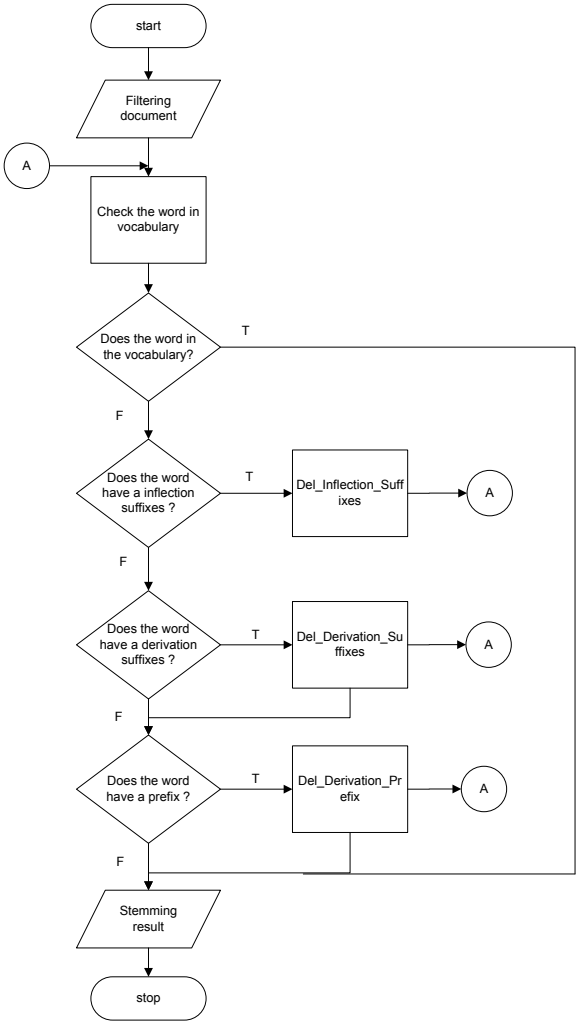

Figure 6 Stemming Algorithm

Process begins with the entry of input filtering results before. Then go into the process of checking the vocabulary. If the word entered is already contained in the vocabulary of the word is to be output directly to the process of stemming, whereas if not, the words is going through the process of checking further. In the program, words that do not qualify in checking vocabulary will undergo three processes, namely:

1. **Delete inflection suffixes** process is words removal process that have the suffix "-lah", "-kah", "-ku", "-mu", or "-nya". for example if there is a word "sebelumnya", in this process the suffix "-nya" in the word "sebelumnya" is removed, so that the results is "sebelum".

2. **Delete derivation suffixes** process is words removal process that have the suffix "-i", "-an" or "-kan". for example if there is the word "pukuli" in this process, the suffix "-i" in the word "pukuli" will be removed, so that the results is "pukul".

3. **Delete prefix derivation** process is words removal process that have the prefix "di-", "ke-", "se-", "te-",

“ber-”, “me-”, or "pe-". for example if there is the word "dibaca", in this process the prefix "di-" in the word "dibaca" will be removed so that the result is "baca". In some words, prefixes can change the form. For example, for the prefix "me-" could turn out to be "mem-", "meng-", "menge-", "menye-", "mempe-", "men-", "meny-", and prefix “pe-“ could turn out to be “per-“, “pem-“, “pen-“, peng-“, “penge-“, “peny-“, pel-“,and else depending on the first letter from the word.

After all words through the process above, the output is stemming results in the form of word stem. For the real example, the input of the filtering process is the word "menyesali". First, system checks whether the word "menyesali" already exists in the database vocabulary. If it is true it will be output directly, but in this case, the word "menyesali", not in the vocabulary database, then the next process is delete inflection suffixes. System check, if the word "menyesali" having the suffix "-lah”, “-kah”, “-ku”, “-mu”, or "-nya". If true, then the word "menyesali" will have the suffix deletion. Yet in the word "meyesali" is no inflection suffix, then process further to delete the derivation suffixes. System checks whether the word "menyesali" having the suffix "-i", "-an" or "-kan". If it is false, then the system will go directly to the next process. Yet in this case, the word "menyesali" there is the suffix "-i", the suffix will undergo a process of elimination. Results obtained from this process in the form of the word "menyesal". Furthermore, the system checks whether the word "menyesal" was in the database, if it is true then the system will go directly to the output. Because it is false, then the process continues to delete prefix derivation. The next process is the delete derivation prefix. System checks whether the word "menyesal" has a prefix. if it is false, the system will immediately to output, but in this case, the word "menyesal" has a prefix, the "me" that change form to "meny-“ (me + sesal = menyesal, according to the Indonesian dictionary), the word “menyesal” having replacement prefix . The prefix "meny-" replaced with vocal alphabets (aiueo) or the letter "s-' that one by one matched to the database vocabulary. Because the word that existing in database is “sesal”, then the output that comes out is the word “sesal”. After that, the process stops.

The example of stemming process in this program can be seen in figure 7.


Tokenizing Filtering Stemming Frekuensi

Filtering	Word Stem
jujur	jujur
ceria	ceria
ramah	ramah
sabar	sabar
humoris	humoris
bergaul	gaul
bertualang	tualang
mengenal	kenal
bersedih	sedih

Figure 7 Stemming Process

3.2 Classification with Naïve Bayes

Naïve bayes method consist of two phases, they are learning phase and classification phase.

1. Learning phase is the phase where the document preprocessing result through the learning process to get a learning data. This process is used to get probabilistic value from P(Vj) and P(Wk|Vj). Flowchart of learning process can be seen in figure 6.

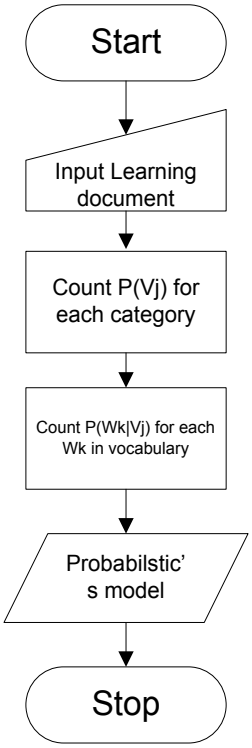

Figure 8 Naïve Bayes Learning Process

The process of learning begins with the input is the learning document then start the forming of vocabulary. Vocabulary is the set of all the unique words of the data training which then the amount being calculate. Furthermore, calculating P (Vj) for each category using the formula:

$$P(V_j) = \frac{|fd(Vj)|}{|D|} \quad (1)$$

Which is fd (Vj) is the number of words in the category j and D is the number of documents used in training. Furthermore, calculating P (Wk | Vj) for each Wk in the vocabulary with formula:

$$P(W_k \mid V_j) = \frac{f(W_k \mid V_j) + 1}{N + |W|} \quad (2)$$

Where P (Wk | Vj) is the amount of occurrences of word wk in the category Vj, N is the amount of all words in

the category Vj and |W| is the number of unique words (distinct) on all training data.

2. The classification phase is the phase where the new document will undergo a process of classification based on data previously coached there. Flowchart for the classification phase can be seen in the Figure 7.

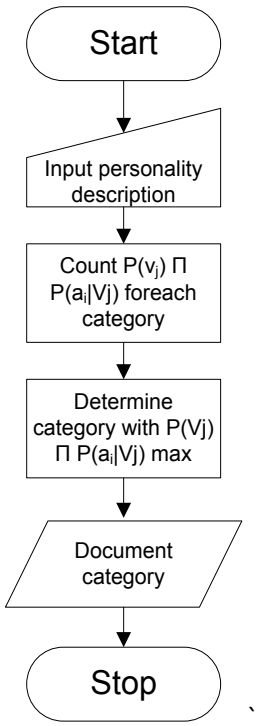

Figure 9 Naïve Bayes Classification Process

In the classification process, the input is personality documents and probabilistic model that has generated in the learning phase. The next stage VMap calculated by the formula:

$$V_{MAP} = \underset{V_j \in V}{\arg\max} P(V_j) \prod_i P(Wk|V_j) \quad (3)$$

After obtained the calculation for each category, then selected categories with maximum VMap that used to classify the personality document. Personality document will be classified according to the categories that have the maximum VMap value.

3.3 Personality Types

According to a book written by Florence Littauer called Personality Plus, more than 400 years before Christ, Hippocrates, a physician and philosopher from Greece, suggested a theory of personality that says that there are basically four types of temperament, they are Sanguine, Choleric, Melancholic and Phlegmatic. Each personality based on Hippocrates theory formed by the bile. Then Galenus refine this theory by stating that the four liquid is present in the body in a certain proportion, whereby if one fluid is more dominant than the other liquids, the liquid can form a personality. Here are the personality types and their characteristics:

1. Sanguine has a cheery and light hearted personality traits, friendly, talkative, likes to smile, outgoing, personality type who would rather party.
2. Choleric personality characterized by a life of passion, hard, heart-flammable, great fighting spirit, optimistic, tough, irritable, regulators, authorities, vengeful, and serious.
3. Personality traits of melancholy have easily disappointed, small guts, grim, pessimistic, fearful, and stiff.
4. Personality Phlegmatic characterized dislike to rush, calm, not easily influenced, loyal, cool, peaceful, relaxed and patient.

In addition there are four mix personalities where there are two dominant types of the same personality. The personality mixture is:

1. Natural mixed personality is the mixed personality that has similar properties. Included are sanguine-Choleric and melancholy-Phlegmatic
2. Complementary mixed personality is the mixed personalities who blend the two are complementary. Included are Choleric-melancholic and sanguine-Phlegmatic
3. Opposite mixed personality is the mixed personality which is the two personality are contradictory. Included are sanguine-melancholic and Choleric-Phlegmatic.

3.4 Couple Compatibility by Type Personality

Everything will attract the opposite. In the personality's type, when there are two types of personalities met will find a match with one another. The cheerful sanguine will improve the life's spirit of melancholy as well as melancholy will make sanguine life more scheduled.

The peaceful phlegmatic dislike to be pressed, but if not, they never find what they want. Meanwhile, choleric is the people who quick to make a decisions, having a goal and diligent, so both of them will match each other.

4. Experiments and Results

Naïve Bayes Method is a supervised learning, so they need require prior knowledge to be able to taking a decision. The success rate of this method depending on initial knowledge that given.

For example, user input the data of personality, as follow: "**Saya adalah orang yang jujur, ceria, ramah, sabar, dan humoris. Saya suka bergaul dengan teman. Saya suka bertualang. Saya suka mengenal hal baru tetapi saya juga sering bersedih**".
That document will through the text mining process, the result will calculate with Naïve Bayes method as seen as table below.

Table 1. Result of Text Mining Process (1)

Category	$P(V_j)$	$P(W_k \mid V_j)$				
		jujur	ceria	ramah	sabar	humoris
Sanguine	1/4	1/200	2/200	2/200	1/200	2/200
Choleric	1/4	1/200	1/200	1/200	1/200	1/200
Melancholic	1/4	1/200	1/200	1/200	1/200	1/200
Phlegmatic	1/4	1/200	1/200	2/200	2/200	1/200

Table 2. Result of Text Mining Process (2)

Category	$P(V_j)$	$P(W_k \mid V_j)$			
		gaul	tualang	kenal	sedih
Sanguine	1/4	1/200	1/200	1/200	1/200
Choleric	1/4	1/200	2/200	1/200	1/200
Melancholic	1/4	1/200	1/200	1/200	1/200
Phlegmatic	1/4	2/200	1/200	1/200	1/200

After knowing the P(Vj) and P(Wk|Vj) then count the VMap for each category.

P(sanguine|document) = $1/4 \times 1/200 \times 2/200 \times 2/200 \times 1/200 \times 1/200 \times 1/200 \times 1/200$

$= 8 / (2{,}048 \, x \, 10^{21})$

$= \mathbf{3{,}09 \times 10^{-21}}$

P(choleric|document) = $1/4 \times 1/200 \times 1/200 \times 1/200 \times 1/200 \times 1/200 \times 2/200 \times 1/200 \times 1/200$

$= 2 / (2{,}048 \, x \, 10^{21})$

$= 0{,}977 \times 10^{-21}$

P(melancholic|document) = $1/4 \times 1/200 \times 1/200 \times 1/200 \times 1/200 \times 1/200 \times 1/200 \times 1/200$

$= 1 / (2{,}048 \, x \, 10^{21})$

$= 0{,}488 \times 10^{-21}$

P(phlegmatic|document) = $1/4 \times 1/200 \times 2/200 \times 2/200 \times 1/200 \times 2/200 \times 1/200 \times 1/200 \times 1/200$

$= 8 / (2{,}048 \, x \, 10^{21})$

$= \mathbf{3{,}09 \times 10^{-21}}$

After see the formula above, the category which has maximum VMap are sanguine and phlegmatic. That's mean the result of text mining with Naïve Bayes Method for the document above is Sanguin and Phlegmatic.

In system, the output of this program are personality types and their potential partner. First, before start using the method to classify the personality, the non registered user must register them. After that, they can login, and use this program.

Figure 10 shows the personality's paragraph which is written in text box area. Besides written the input, user can also input it through the file with .txt extension.

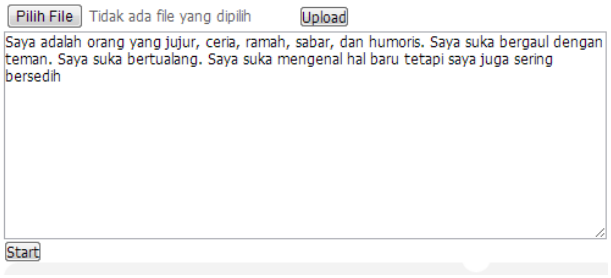

Figure 10 Input personality's data process

After the finish written or upload the data, click the start button. Then the result will appear as shown in figure 10.

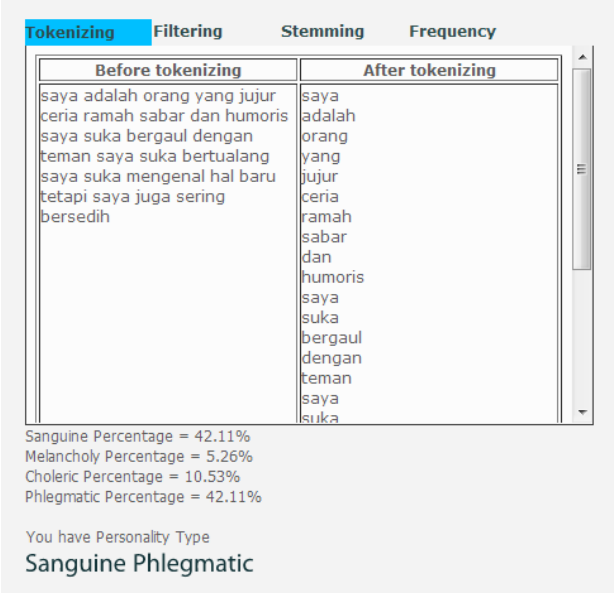

Figure 11 Result of Personality Type

After knowing the personality types, users can find their potential mates. As example above, user has a complementary mixed personality, which is sanguine and phlegmatic. As the theory of couple compatibility, the sanguine is a mate of melancholy and phlegmatic is a mate of choleric. So their mate must be a person who have melancholy, choleric, or two of them. Figure 11 will show the result of the matching couples.

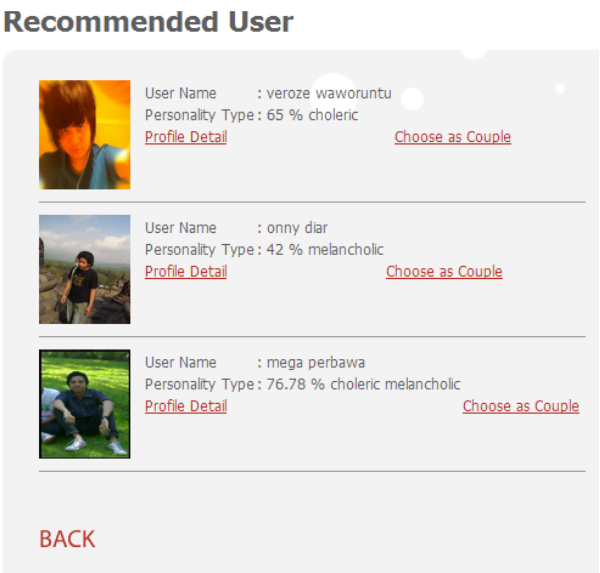

Figure 12 Result of matching couples

This experiment use 40 documents training and has 160 learning documents. In table 3, is the result of the details of personality classification for 40 training data which has been classified. There are 3 errors (error) which has produced the three data are unidentified category. So the percentage error reached,

$$\text{Accuracy percentage} = \frac{sum\ of\ correct\ \ classifica\ tion}{training\ \ documents} \text{ x } 100\% \quad (4)$$

$$= \frac{37}{40} \text{ x } 100\% = 92{,}5\%$$

Table 3. Result of Classification Training Document

Document Number	Classification Result	True/Flase
1	Phlegmatic	True
2	Melancholy	True
3	Phlegmatic	True
4	Phlegmatic	True
5	Sanguine Choleric	True
6	Melancholy	True
7	Phlegmatic	True
8	Phlegmatic	True
9	Sanguine	True
10	Choleric	True
11	Sanguine	True
12	Choleric	True
13	Melancholy	True
14	Melancholy	True
15	Unidentified category	False
16	Unidentified category	False
17	Sanguin Phlegmatic	True
18	Choleric Melancholy	True
19	Sanguine Phlegmatic	True
20	Choleric Phlegmatic	True
21	Sanguine	True
22	Melancholy	True
23	Choleric	True
24	Choleric Phlegmatic	True
25	Sanguine Melancholy	True
26	Melancholy	True
27	Sanguine	True
28	Sanguine	True
29	Phlegmatic	True
30	Sanguine Melancholy	True
31	Choleric melancholy	True
32	Sanguine choleric	True
33	Unidentified category	False
34	Choleric	True
35	Choleric	True
36	Choleric Phlegmatic	True
37	Melancholy	True
38	Sanguine Phlegmatic	True
39	Melancholy	True
40	Sanguine Phlegmatic	True

With the number of training data with error percentage as such, the 40 training data will use as learning data in the database for classify the training data in subsequent experiments and is expected to shrink error percentage in selecting or classifying personality types.

5. Conclusion

This experiment has successfully obtained the type of personality and finds a mate based on personality types by using the text mining with Naïve Bayes method for personality classification. In this experiment, some of the user data personality is used as learning document in the learning process of Naive Bayes methods. The success rate of the classification depends on the amount of learning document used. Personality classification process is done by the determination of the biggest VMap from each category. For matching couple output, the programs use Personality compatibilities theory, where the matching couples are the couples who have opposite personalities.

Acknowledgments

Our thank goes to Department of Information Technology Udayana University, Bali, who has helped organize this research's in Indonesia.

References

[1] Reddy V, Siva RamaKrishna, dkk. Classification of Movie Reviews Using Complemented Naïve Bayesian Classifier: Prithyi Information Solutions Limited: India

[2] Hamzah, Amir. Text Classification with Naive Bayes classifier (NBC) for Abstract Grouping Text and Academic News. Prosidign Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III. Yogyakarta. 3 November 2012.

[3] Abdur Rozaq, Abdur., Agus Zainal Arifin., Diana Purwitasari. Arabic Language Text Document Classification using Naive Bayes Algorithm: Surabaya

[4] Kim, Jangwoo., Daniel X. Le, and George R. Thoma. Naïve Bayes Classifier for Extracting Bibliographic Information from Biomedical Online Articles: National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894: USA

[5] Rothbart, Mary.K., Stephan A. Ahadi., David E. Evans. Temperament and Personality: Origins and Outcomes. Journal of Personality dan Social Psychology 2000, Vol. 78. No 1. 122-135

[6] Littauer, Florence. 1992. Personality Plus. Jakarta Barat: Binarupa Aksara

[7] Aprilia, Krisma Dini. 2008 Application of Naive Bayes for classification SMS Customer's Voice (Case Study PT. Pertamina UPMS V Surabaya): Stikom Digilib : Surabaya

[8] Saraswati, Ni Wayan Sumartini. Text Mining dengan Metode Naïve Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis: Denpasar. 2011

[9] Maharsi, Lisa. Text Document Keywords Extraction Using Naïve Bayes Method: Bandung. 2009

[10] Anugroho, Prasetyo., Idris Winarno., S.ST M.Kom., Nur Rosyid M., S.Kom. Spam Email Classification with Naïve Bayes Classifier Method use Java Programming: Surabaya.

[11] Indranandita, Amalia., Budi Susanto, and Antonius Rachmat C. Classification System and Journal Search using Naive Bayes Methods and Vector Space Model. Jurnal Informatika, Volume 4 Nomor 2, November 2008.

[12] Trisedya, Bayu Distiawa and Hardinal Jais. Document Classification using Naive Bayes algorithm with the addition of Parameter Probability Parent Category: Jakarta.2009

[13] Nurhayati, Sri. Implementation of Text Mining for Classification of Traditional Arts with NBC method (Naive Bayes Classifier): Bandung

[14] Destuardi.I dan Surya Sumpeno. 2009 Emotion Classification for Indonesian Language Text Using Naïve Bayes Method: Jurnal Teknik Elektro ITS : Surabaya

[15] Feldman, Ronen., James Sanger. 2007. The Text Mining Handbook. United Kingdom: Cambridge University Press

[16] Khodra, Masayu Leylia. Text Mining Text Categorization Naïve Bayes : Informatika ITB: Bandung

AA Ketut Agung Cahyawan, ST., MT received his ST degree and MT degree in Electrical Engineering from Institut Teknologi Bandung. He is lecturer at Electrical Engineering Department (major in Computer System and Informatics) of Udayana University, lecturer at Information Technology Department of Udayana University

Ni Made Ari Lestari study in Information Technology, Department of Information Technology Udayana University since August 2008, and now working her research for S.Ti. degree in Information Technology.

Dr. I Ketut Gede Darma Putra, S.Kom., MT received his S.Kom degree in Informatics Engineering from Institut Teknologi Sepuluh Nopember University, his MT. degree in Electrical Engineering from Gajah Mada University and his Dr. degree in Electrical Engineering from Gajah Mada University. He is lecturer at Electrical Engineering Department (major in Computer System and Informatics) of Udayana University, lecturer at Information Technology Department of Udayana University.