Analysis Of Data Mining Methods Naive Bayes Classifier (NBC)

Yuliana[1], Erlangga[1]
[1] Information System, Computer Science Faculty, Bandar Lampung University, Indonesia

1. Introduction

1.1 Background

At present, the development of the information that is unbelievably experienced rapid development is no exception to the application database. In a database application is needed to store important data and the development of increasingly large storage media so that data can be stored on a database of many in the data storage media. Indirectly lead to the accumulation of data and unwittingly data it stores a lot of information that can be utilized. Therefore, the necessary data mining, which is a concept to discover knowledge hidden within a database. Data mining is a semi-automatic process that uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify potential knowledge and useful information that is stored in large databases. Find information hidden within a large data one method that can be used is a method of Naive Bayes classifier (NBC). Methods Naive Bayes classifier (NBC) is one method that is likely a simple data classification based on Bayes theorem, assuming between the explanatory variables are independent (independent). This method is one of many methods used to classification hidden data and get information from a lot of data. Naive Bayes classifier method assumes that the presence or absence of a particular occurrence of a group is not associated with the presence or absence of other events. Naive Bayes classifier method can be combined with a Decision Support System for the method does not require weights to do the calculations, but only using the probability data that already exists. 2 Reasons for choosing Naive Bayes classifier method in this study because of the classification of the opportunities that simple and easy to understand, efficient computation, and the results can be recognized more accurately than other methods, and has the ability to classification by category in a simple mathematical form. Thus it is expected that this method will be able to find information hidden from much of the data with the classification by category.

2. Basic Theory

2.1 Literature Review

The review of the literature used by the authors are as follows: a. Classification and Academic Reference Book Search Methods Using Naive Bayes classifier (NBC) Mulawarman Informatics Journal Vol.10 No.1 February 2013. The author Agus Setiawan, et al, from the Department of Computer Science, Science Faculty, Mulawarman. The design of the classification and search books using use case diagrams and activity diagrams, and research methods Naive Bayes classifier produce convenience of visitors in finding the required book and facilitate library employees in managing existing books in the library. b. Classification of Radar Malang Local News Naive Bayes Method Using N-Gram Features Journal of Scientific Technology and Information ASIA (JITIKA) Vol.10 No.1 February 2016. Author Denny Nathaniel Chandra, et al, from the Department of Computer Science, Graduate School, University of Education Ganesha. Classification of news using the N-gram model building, and Naive Bayes classifier method produces an effective and accurate news with accuracy maximum value of 78.66%. c. Implementation Of Naive Bayes Classification Method To Predict Graduation Time Of IBI Dharmajaya Scholar International Conference On Information Technology And Business ISSN 2460-7223. Author Ketut Artaye, Informatics Engineering IBI Dharmajaya Lampung. Predicting the time of graduation with data-driven student who had graduated a few years ago, using Naive Bayes classifier predictions graduation coming year and maintain the quality of university graduation. 4 d. Personality Types Indonesian Classification for Text in Partners SearchingSAT Website Using Naïve Bayes Methods, International Journal of Computer Science Issues (IJCSI) Vol.10, Issues 1, 3, January 2013. ISSN (Print): 1694-0784 | ISSN (Online): 1694-0814. Writer Ni Made Ari Lestari, et al, from the Department Of Information Techonology, Udaya University. Classification of personality types using the Naive Bayes classifier produce ease in knowing the personality of themselves and can find a partner with the same personality or complementary.

3. Results Analysis and Discussion

In this study the authors analyzed four different journals with a case that includes the results anddiscussion using the Naive Bayes classifier. Here are the results and discussion journals have been the author of the analysis of the four journals as follows. 3.1 Analysis of Results The result has been the author analyzes in the four journals as follows: a. Journal of Classification and Academic Reference Book Search Methods Using Naive Bayes classifier (NBC) This journal has a case such as students or visitors is difficult to get the academic reference books in the library, because banyanya books in the library to make visitors difficult to find a book that is needed. Solutions to allow visitors to search for books in the library by category, namely with data mining uses Naive Bayes classifier method. The data used in this paper that the data train coming from the online catalog of various book publishers Indonesian, total books used in the training data is numbered 250 books, for the classification comes from a collection of books found in libraries and test data used were 150 book. Classified into five categories: computer programming, computer networks, databases, multimedia, and operating systems. The following discussion of the calculation process Naive Bayes classifier in this journal. Step 1 text classification stages, this stage is the beginning of a process input text data and produce the output of the pattern as a result of interpretation (Events and Zohar, 2002). As illustrated in the stages of text categorization can be seen in figure 1. 1(Suhartono, 2013).

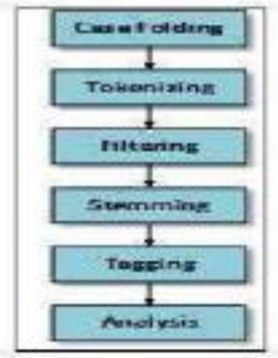


Figure 1. Classification Stages text

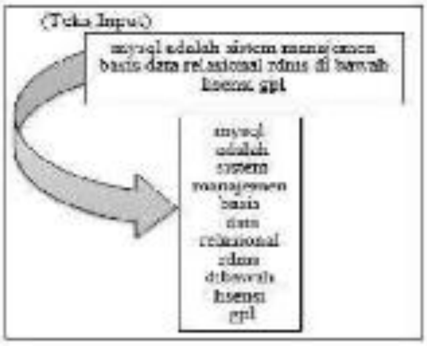


Figure 2. Process Tokenizing

Step 2 initial stages of text classification is text preprocessing to prepare the text into data that will undergo treatment at a later stage, some of the actions taken at this stage are: stages case folding and tokenizing stages, these stages can be seen in Figures 2 and 3.

Tokenizing process is to select the contents of the text so that a single word. Step 3 phase transformations text, is made to reduce the number of words contained by eliminating stopword and change the wording in its basic form (stemming). Stopword removal process is used to reduce the workload of the system, the filtering stage which is a stage which took the important words from the token. Can be seen in Figure 4.13

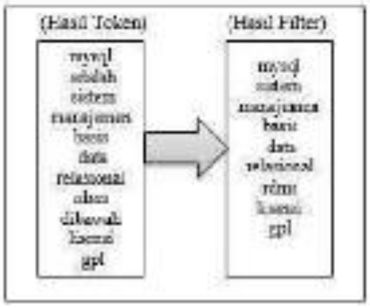


Figure 3. Filtering Process

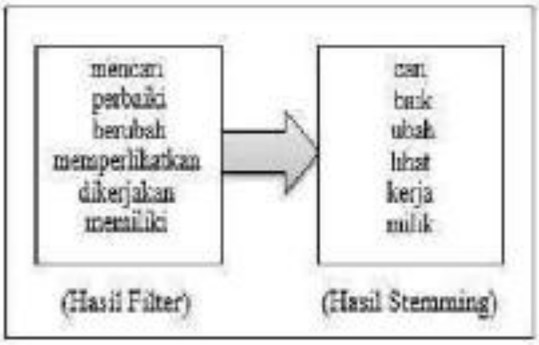


Figure 4. Process Stemming

Stemming Further stages to reduce the word into its basic form or in other words, is a process that provides a mapping between different words with different morphologies into basic shape (stem). This phase is to find the root (root) words of each word filter results. can be seen in Figure 3. Endangered5 stages included in the method is the method of Naive Bayes classifier which is the method used to classify blocks of text, this algorithm utilizes the methods of probability and statistics, to predict the probability of the future based on the experience of earlier (Ismail, 2008).Classification Naive Bayes classifier is done by finding the probability, the probability category if known text.At the time of classification, Bayes approach will result in the highest category label probabilitywith input attributeswhere a1 is the first word, and so on. a probability value computed values of and .So the training text for Naive Bayes algorithm can be equation definition and the probabilitysaid for each category calculated at the time of training. where | is the amount of data on category and is data used in training. While is the number of occurrences on category, is the sum of all the wordscategories and is the number of words in all training data. Based on the equation phases of the training process and the classification.

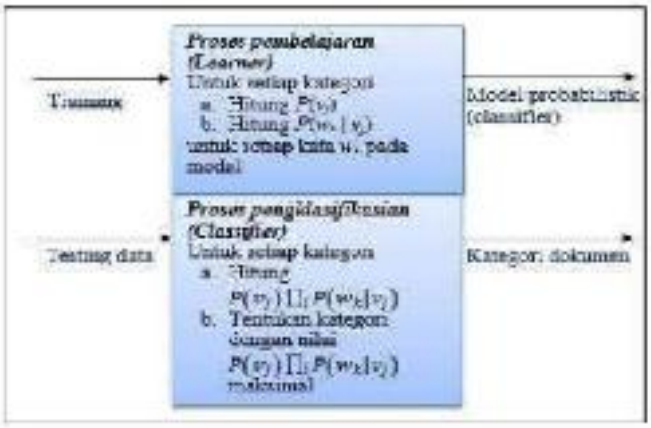
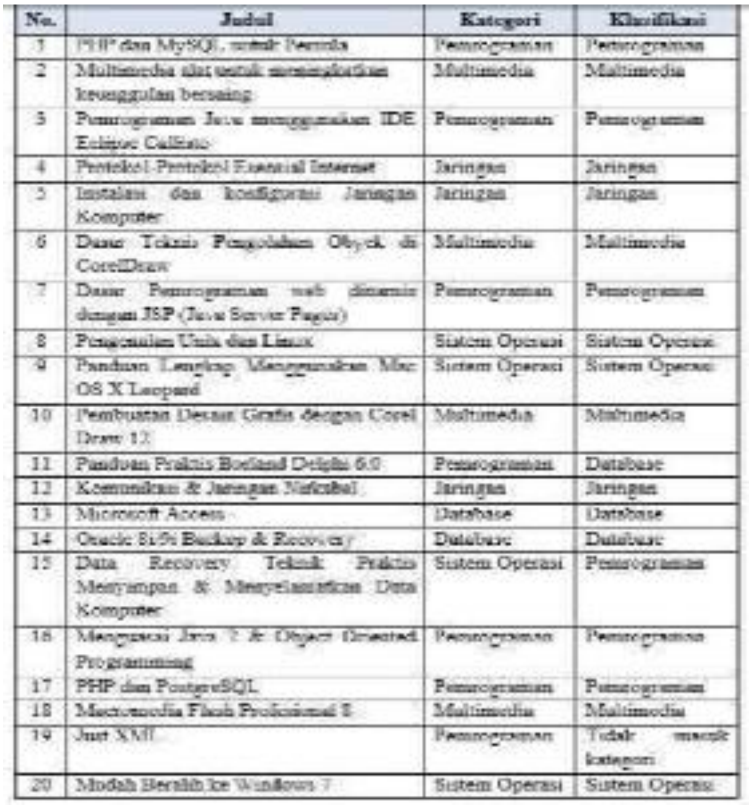

Training
Proses pembelajaran (Learner)
Untuk setiap kategori
a. Hitung P(vj)
b. Hitung P(wk | vj)
untuk setiap kata wk pada model
Model probabilistic (classifier)
Testing data
Proses pengklasifikasian (Classifier)
Untuk setiap kategori
a. Hitung P(vj) ∏i P(wk|vj)
b. Tentukan kategori dengan nilai P(vj) ∏i P(wk|vj) maksimal
Kategori dokumen

Figure 5. Stages of Document Classification Algorithm Naive Bayes

Step 6 of this journal have test data from a collection of books contained in the library of East Kalimantan Province. The test data used were 150 books. The following table 1 the results of the classification of books.

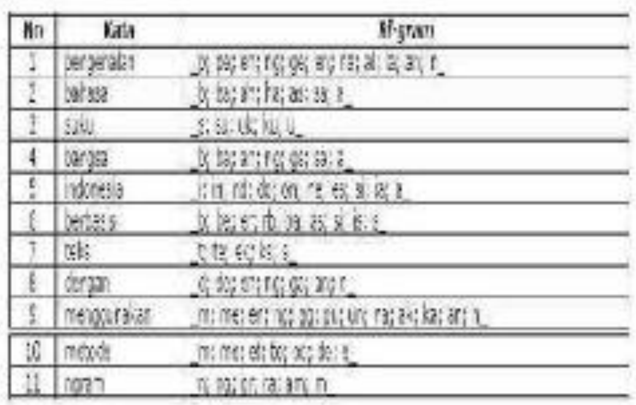
Table 1. Results of the Library Book KalTim classification



No.	Judul	Kategori	Klasifikasi
1	PHP dan MySQL untuk Pemula	Pemrograman	Pemrograman
2	Multimedia alat untuk meningkatkan keunggulan bersaing	Multimedia	Multimedia
3	Pemrograman Java menggunakan IDE Eclipse Callisto	Pemrograman	Pemrograman
4	Protokol-Protokol Esensial Internet	Jaringan	Jaringan
5	Instalasi dan konfigurasi Jaringan Komputer	Jaringan	Jaringan
6	Dasar Teknis Pengolahan Obyek di CorelDraw	Multimedia	Multimedia
7	Dasar Pemrograman web dinamis dengan JSP (Java Server Pages)	Pemrograman	Pemrograman
8	Pengenalan Unix dan Linux	Sistem Operasi	Sistem Operasi
9	Panduan Lengkap Menggunakan Mac OS X Leopard	Sistem Operasi	Sistem Operasi
10	Pembuatan Desain Grafis dengan Corel Draw 12	Multimedia	Multimedia
11	Panduan Praktis Borland Delphi 6.0	Pemrograman	Database
12	Komunikasi & Jaringan Nirkabel	Jaringan	Jaringan
13	Microsoft Access	Database	Database
14	Oracle 8i/9i Backup & Recovery	Database	Database
15	Data Recovery Teknik Praktis Menyimpan & Menyelamatkan Data Komputer	Sistem Operasi	Pemrograman
16	Menguasai Java 2 & Object Oriented Programming	Pemrograman	Pemrograman
17	PHP dan PostgreSQL	Pemrograman	Pemrograman
18	Macromedia Flash Profesional 8	Multimedia	Multimedia
19	Just XML	Pemrograman	Tidak masuk kategori
20	Mudah Beralih ke Windows 7	Sistem Operasi	Sistem Operasi

classification results indicating "not categorized", explaining that the results stemming words contained in the title is neither contained in the training database. 16 Application of Naive Bayes classifier (NBC) are also implemented when the book has been successfully uploaded, the system will automatically perform a classification to determine the category of books. So administrators no longer need to enter the book category. if the title is entered is already contained in the database will display a notification that the title is already there so can not be added again. b. Classification Journal Local News Radar Malang Naive Bayes Method Using N-Gram Features This journal has a case newsreader should be more careful in getting the relevant information, in accordance with what is desired, and can easily find news information about a particular event. Because the information is experiencing a large increase, making the volume of electronic news Indonesian language increasingly large, valuable resources, and allows multiple users to alter information, multiply and produce new information. Solution to classify documents using the word Naive Bayes classifier method with the features of the N-gram. Data or documents used in this paper are drawn from the training data www.kompas.com and testing of data taken from www.radarmalang.co.id. Classified into seven categories, namely: economic, news, education, health, sports, entertainment, and others. Here are the results of the stages and processes for classification with Naive Bayes classifier calculations in this journal. Step 1 Text Mining This stage is the process of extracting a pattern in the form of useful information and knowledge from large amounts of text data sources, such as word documents, pdf, and more. Text mining at this stage that begins with apply structure of the source text data and the continued phase information extraction and knowledge of relevant data structured text is by using techniques and tools 17 together with data mining, the process is carried out mining the text of which is summary automatic , categories of documents, assignment text, and more. Step 2 preprocessing, the text preprocessing step, there are several stages such as stages of folding case is to convert all the letters in the document to lowercase. Once all the letters into small stages next. Stages tokenizing which separates the rows of words in sentences, paragraphs or pages into a token or a piece of a single word (termed word), after the word into pieces next word Stages filtering that stage took the important words from the token, can using algorithms stoplist (waste words are less important) or wordlist (save important words), such as "the", "and", "in", and others. The next stage is to look for the root stages stemming words / phrases from the filtering step is conducted the process of the various forms of a word into a representation of the same. Step 3 Features N-gram is a method used to pick up the pieces of the different character number n of a word that continuity is read from the source text to the end of the document, to assist in taking the pieces of words in the form of lower-case characters then do padding with blanks beginning and end of a word. Examples of the word "TEXT" can be broken down into several n-gram below (represent blank): Uni-gram: T, E, X, T Bi-gram: _T, TE, EX, XT, T_ Tri-gram: _TE , TEX, EXT, XT_ 18 quard gram: _TEX, TEXT, EXT_ Quintgram: _TEXT, TEXT Phase Formation of N-gram model of a document, based on the frequency gram that appear in the document, the frequency for gram be plus one, if not gram it will be added into the table a number of times, the document gram model building using the bi-gram in a document that contains only one sentence, "the introduction of language-based Indonesian tribe using gram". Will produce a bi-gram (2-gram) in Table 3.

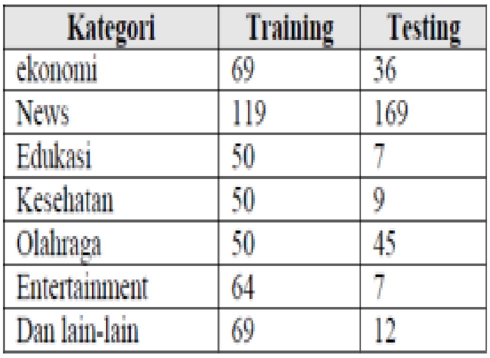
Table 2. Example Formation of N-gram.



No	Kata	N-gram
1	pengenalan	_p; pe; en; ng; ge; en; na; al; la; an; n_
2	bahasa	_b; ba; ah; ha; as; sa; a_
3	suku	_s; su; uk; ku; u_
4	bangsa	_b; ba; an; ng; gs; sa; a_
5	indonesia	_i; in; nd; do; on; ne; es; si; ia; a_
6	berbasis	_b; be; er; rb; ba; as; si; is; s_
7	teks	_t; te; ek; ks; s_
8	dengan	_d; de; en; ng; ga; an; n_
9	menggunakan	_m; me; en; ng; gg; gu; un; na; ak; ka; an; n_
10	metode	_m; me; et; to; od; de; e_
11	ngram	_n; ng; gr; ra; am; m_

Stages Classifier Naive Bayes method, the method of classification by using methods of probability and statistics is to predict future opportunities based on past experience. The basis of the theorem Naive Bayes Classifier used in programming is a formula Bayes following: (|)=((|)* ( ))/ ( )Opportunity events as set of opportunities ,opportunities, and opportunities .In its application later the formula changed to: (|)=((|) * ( ) / ( )19 Naive Bayes classifier is a model simplification of algorithms Bayes matches in the classification of text or documents. The equation is: = (1, 2 ... |) ( ) (1, 2 ...) (1, 2 ...)It is a constant, so it is removed into = (1, 2 ... |) ( ) Because (1, 2 ... |) ( ) is difficult to quantify, it will be assumed that every word in the document has no relevance. = ( ) Π) |(Description: ( )= | | | |)|(= + 1 + | | Where to: ( ):is the probability of each document to a collection of documents. (|):is the probability of occurrence of the word on a document withclass categories.| | : Is the frequency of documents in each category. | | : Is the number of documents. : is frequency words in each category is the number of words in the test documents. 20 is the frequency of word in each category is the number of words in the test documents. Experiment 1 The first test conducted randomly selecting training data and test data of 764 news data. The selected training data totaled 471 news data from each of the categories, and the remaining 293 news data were used as test data. Below is the composition of the distribution of test data and training data for this experiment.

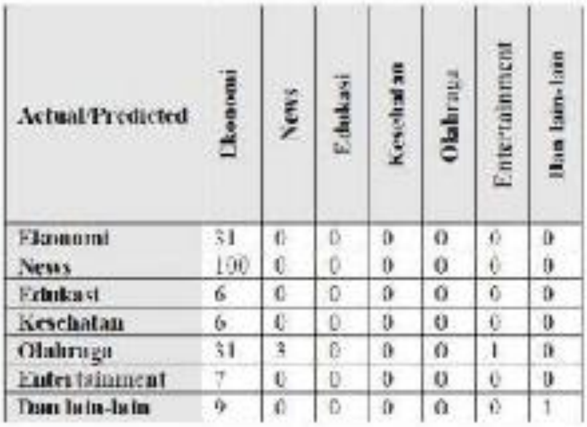
Table 3. Train and Test Data Composition Experiment 1



Kategori	Training	Testing
ekonomi	69	36
News	119	169
Edukasi	50	7
Kesehatan	50	9
Olahraga	50	45
Entertainment	64	7
Dan lain-lain	69	12

Is obtained through experiment or accuracy prediction of 78.66% (accurately predicted as many as 601 out of 764 news data test).And the error of 14.80% (less precisely predicted many as 163 out of 764 news data test). Table confusion matrix of tests conducted.

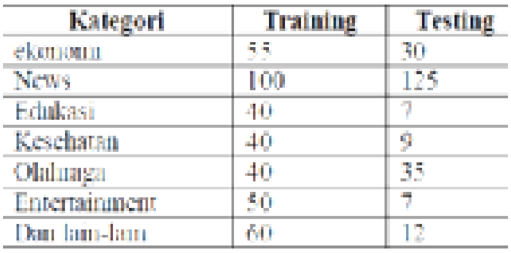
Table 4. Experiment 1 Confusion Matrix



Actual/Predicted	Ekonomi	News	Edukasi	Kesehatan	Olahraga	Entertainment	Dan lain-lain
Ekonomi	31	0	0	0	0	0	0
News	100	0	0	0	0	0	0
Edukasi	6	0	0	0	0	0	0
Kesehatan	6	0	0	0	0	0	0
Olahraga	31	3	0	0	0	1	0
Entertainment	7	0	0	0	0	0	0
Dan lain-lain	9	0	0	0	0	0	1

Experiment 2 Experiments both know the amount of data 610 with the seven categories of news data, as many as 385 news training data and test data as much as 225 news, the composition of the distribution of training data and test data as in table 6 table 3.

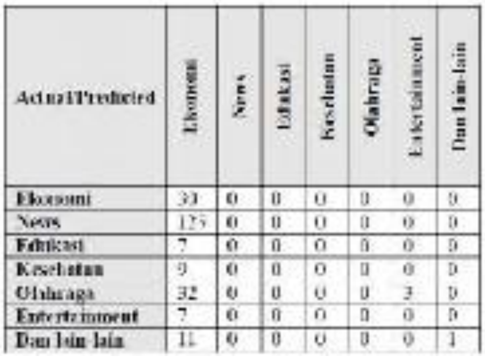
Table 5. Composition of training data and test data in experiment 2



Kategori	Training	Testing
ekonomi	55	30
News	100	125
Edukasi	40	7
Kesehatan	40	9
Olahraga	40	35
Entertainment	50	7
Dan lain-lain	60	12

experiment is obtainable accuracy of the prediction accuracy of 68.20% as much as 416 news accurately and error amounted to 31.80% as much as 194 news. The trial metrix confusion table 2.

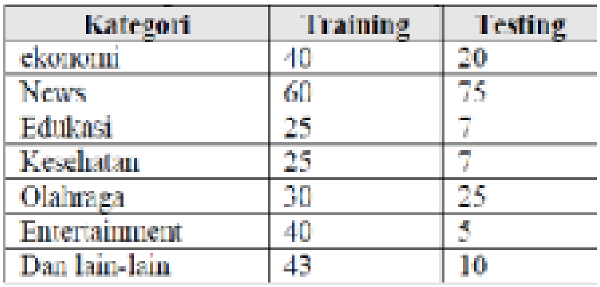
Table 6. Confusion metrix experiment 2



Actual\Predicted	Ekonomi	News	Edukasi	Kesehatan	Olahraga	Entertainment	Dan lain-lain
Ekonomi	33	0	0	0	0	0	0
News	125	0	0	0	0	0	0
Edukasi	7	0	0	0	0	0	0
Kesehatan	9	0	0	0	0	0	0
Olahraga	32	0	0	0	0	3	0
Entertainment	7	0	0	0	0	0	0
Dan lain-lain	11	0	0	0	0	0	1

Experiment 3 Experiments carried out by the data number 314, with as many as 180 training data and test data as much as 134, with seven categories, the composition of the distribution of test data and training data for this experiment can be seen in table 4. 22

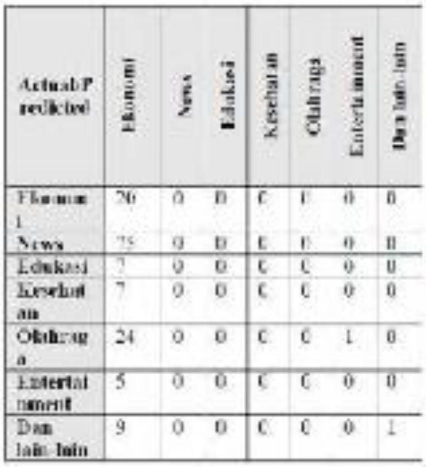
Table 7. Composition of training data and test data in experiment 3



Kategori	Training	Testing
ekonomi	40	20
News	60	75
Edukasi	25	7
Kesehatan	25	7
Olahraga	30	25
Entertainment	40	5
Dan lain-lain	43	10

Through this experiment obtained prediction accuracy or accuracy of 59.24% or as much as 186 news accurately predicted, and the error amounted to 40.76% or as much as 128 error message. Confusion matrix of the tests conducted are shown in Table 8.

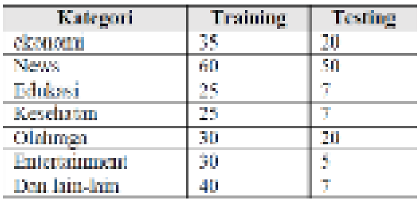
Table 8. Confusion matrix experiment 3



Actual\Predicted	Ekonomi	News	Edukasi	Kesehatan	Olahraga	Entertainment	Dan lain-lain
Ekonomi	20	0	0	0	0	0	0
News	75	0	0	0	0	0	0
Edukasi	7	0	0	0	0	0	0
Kesehatan	7	0	0	0	0	0	0
Olahraga	24	0	0	0	0	1	0
Entertainment	5	0	0	0	0	0	0
Dan lain-lain	9	0	0	0	0	0	1

Trial 4 Trial fourth the amount of data that is shared news data 361 to 245 are used as training data and 116 are used as the test data, test data and the composition division training data in the experiments shown in table 9. 23

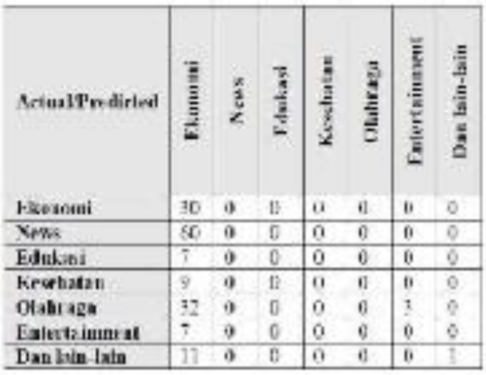
Table 9. Composition of the division of training data and test data in 4 experimental



Kategori	Training	Testing
ekonomi	35	20
News	60	50
Edukasi	25	7
Kesehatan	25	7
Olahraga	30	20
Entertainment	30	5
Dan lain-lain	40	7

prediction accuracy obtained through experiment or accuracy of 63.93% or as much as 238 news data that is accurate, and error of 34.07% or as much as 123 Data news errors. The following table 10 is confusion matrix of experiments performed.

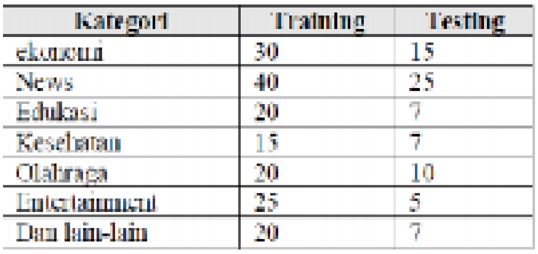
Table 10. Confusion matrix Experiment 4



Actual/Predicted	Ekonomi	News	Edukasi	Kesehatan	Olahraga	Entertainment	Dan lain-lain
Ekonomi	30	0	0	0	0	0	0
News	50	0	0	0	0	0	0
Edukasi	7	0	0	0	0	0	0
Kesehatan	9	0	0	0	0	0	0
Olahraga	32	0	0	0	0	3	0
Entertainment	7	0	0	0	0	0	0
Dan lain-lain	11	0	0	0	0	0	1

Experiment 5 fifth Test with the amount of data that is shared news data 264 to 170 are used as training data and 76 are used as the test data, the composition of the distribution of test data and training data in the experiments shown in Table 11. 24

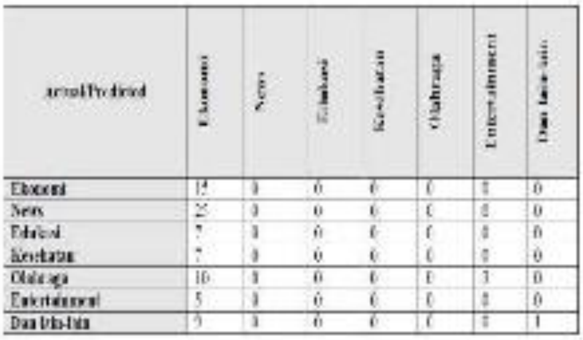
Table 11. the composition distribution of training data and test data in the trial5



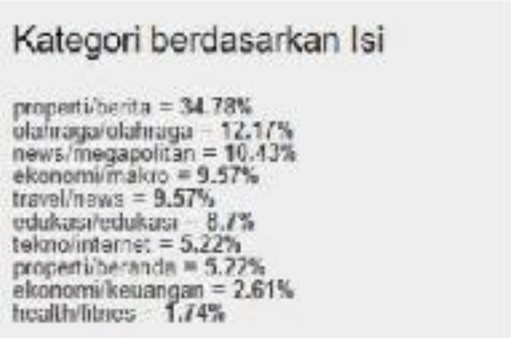
Kategori	Training	Testing
ekonomi	30	15
News	40	25
Edukasi	20	7
Kesehatan	15	7
Olahraga	20	10
Entertainment	25	5
Dan lain-lain	20	7

GainedThrough experiments or accuracy of prediction accuracy of 74.39% or as much as 186 news data that is accurate, and error of 23.61% or as much as 63 news data errors. The following table 12 is confusion matrix of experiments performed.

Table 12. Confusion matrix experiments



actual/Predicted	Ekonomi	News	Edukasi	Kesehatan	Olahraga	Entertainment	Dan lain-lain
Ekonomi	15	0	0	0	0	0	0
News	25	0	0	0	0	0	0
Edukasi	7	0	0	0	0	0	0
Kesehatan	7	0	0	0	0	0	0
Olahraga	10	0	0	0	0	3	0
Entertainment	5	0	0	0	0	0	0
Dan lain-lain	5	0	0	0	0	0	1



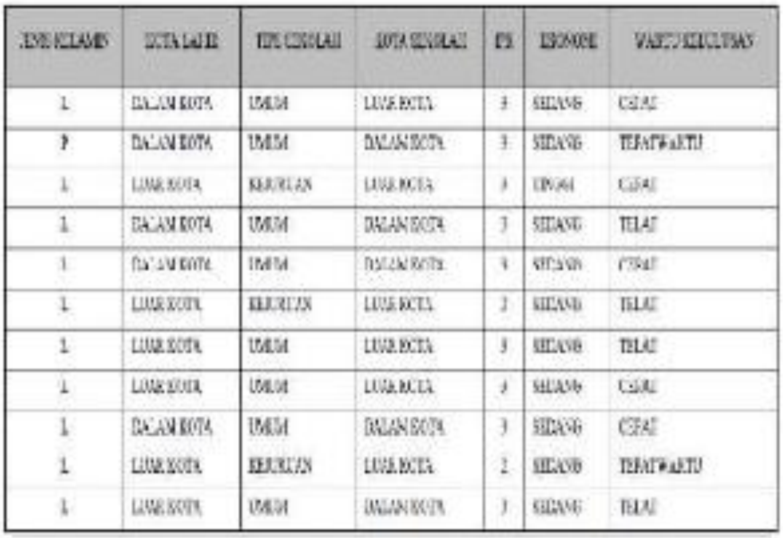
Kategori berdasarkan Isi

properti/berita = 34.78%
olahraga/olahraga = 12.17%
news/megapolitan = 10.43%
ekonomi/makro = 9.57%
travel/news = 9.57%
edukasi/edukasi = 8.7%
tekno/internet = 5.22%
properti/beranda = 5.22%
ekonomi/keuangan = 2.61%
health/fitnes = 1.74%

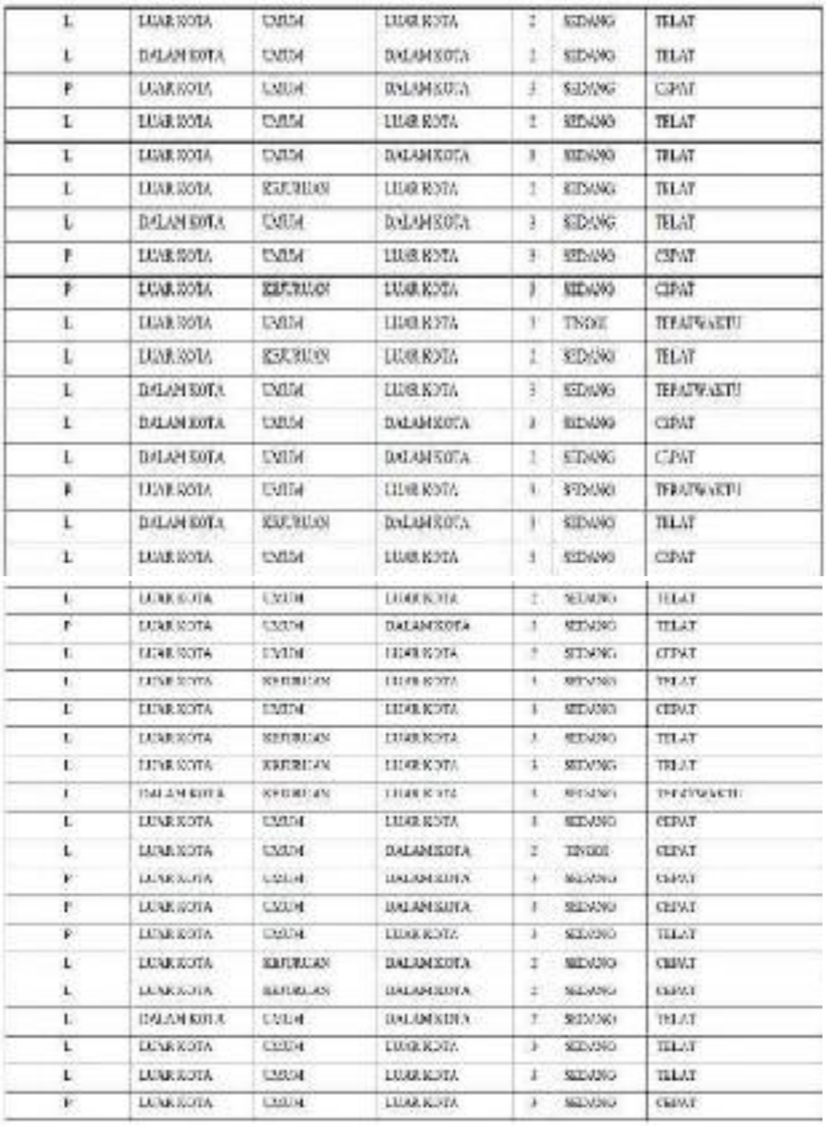
Figure 6. Results Categorizing news

Evaluation Trial Results calculated value with Naive Bayes classification accuracy obtained from the five trials in a row is 78.66%, 68.20%, 59.24%, 63.93% and 74.39%. the value of the lowest accuracy was 59.24% and the highest accuracy is 78.66%. Through further investigation, the possibility of the accuracy can be improved if the training data obtained more and more, because the naïve Bayes classification is a supervised learning method that is highly dependent on the training data. c. Journal Implementation Of Naive Bayes Classification Method To Predict Graduation Time Of IBI Dharmajaya Scholar This journal has a case that universities retain quality views of the average time to graduation for a good university and a lot of public interest is a university with graduation period of time and have quality good. Solutions to maintain the quality of the graduation period is to predict the future based on previous events or past events using the method Naive Bayes classifier. The data used in this journal test data taken from students who had graduated in 2011-2012, from 191 the data students, 50 records were taken as training data. Classified into three categories namely: fast, timely, and slow or late. Following the discussion on this journal, which is to predict the time of graduation. Step 1 testing process, data is divided into two parts, training and test algorithms of data, training data to use to create a probability table while the test data used to test the probability table. The training data table 13 26

Table 13. Data Training



JENIS KELAMIN	KOTA LAHIR	TIPE SEKOLAH	KOTA SEKOLAH	IPS	EKONOMI	WAKTU KELULUSAN
1	DALAM KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT
2	DALAM KOTA	UMUM	DALAM KOTA	3	SEDANG	TEPAT WAKTU
1	LUAR KOTA	KEJURUAN	LUAR KOTA	3	TINGGI	CEPAT
1	DALAM KOTA	UMUM	DALAM KOTA	3	SEDANG	TELAT
1	DALAM KOTA	UMUM	DALAM KOTA	3	SEDANG	CEPAT
1	LUAR KOTA	KEJURUAN	LUAR KOTA	3	SEDANG	TELAT
1	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	TELAT
1	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT
1	DALAM KOTA	UMUM	DALAM KOTA	3	SEDANG	CEPAT
1	LUAR KOTA	KEJURUAN	LUAR KOTA	2	SEDANG	TEPAT WAKTU
1	LUAR KOTA	UMUM	DALAM KOTA	3	SEDANG	TELAT



L	LUAR KOTA	UMUM	LUAR KOTA	2	SEDANG	TELAT
L	DALAM KOTA	UMUM	DALAM KOTA	2	SEDANG	TELAT
P	LUAR KOTA	UMUM	DALAM KOTA	3	SEDANG	CEPAT
L	LUAR KOTA	UMUM	LUAR KOTA	2	SEDANG	TELAT
L	LUAR KOTA	UMUM	DALAM KOTA	3	SEDANG	TELAT
L	LUAR KOTA	KEJURUAN	LUAR KOTA	2	SEDANG	TELAT
L	DALAM KOTA	UMUM	DALAM KOTA	3	SEDANG	TELAT
P	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT
P	LUAR KOTA	KEJURUAN	LUAR KOTA	3	SEDANG	CEPAT
L	LUAR KOTA	UMUM	LUAR KOTA	3	TINGGI	TEPAT WAKTU
L	LUAR KOTA	KEJURUAN	LUAR KOTA	2	SEDANG	TELAT
L	DALAM KOTA	UMUM	LUAR KOTA	3	SEDANG	TEPAT WAKTU
L	DALAM KOTA	UMUM	DALAM KOTA	3	SEDANG	CEPAT
L	DALAM KOTA	UMUM	DALAM KOTA	2	SEDANG	CEPAT
P	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	TEPAT WAKTU
L	DALAM KOTA	KEJURUAN	DALAM KOTA	3	SEDANG	TELAT
L	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT
L	LUAR KOTA	UMUM	LUAR KOTA	2	SEDANG	TELAT
P	LUAR KOTA	UMUM	DALAM KOTA	3	SEDANG	TELAT
L	LUAR KOTA	UMUM	LUAR KOTA	2	SEDANG	CEPAT
L	LUAR KOTA	KEJURUAN	LUAR KOTA	3	SEDANG	TELAT
L	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT
L	LUAR KOTA	KEJURUAN	LUAR KOTA	3	SEDANG	TELAT
L	LUAR KOTA	KEJURUAN	LUAR KOTA	3	SEDANG	TELAT
L	DALAM KOTA	KEJURUAN	LUAR KOTA	3	SEDANG	TEPAT WAKTU
L	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT
L	LUAR KOTA	UMUM	DALAM KOTA	2	TINGGI	CEPAT
P	LUAR KOTA	UMUM	DALAM KOTA	3	SEDANG	CEPAT
P	LUAR KOTA	UMUM	DALAM KOTA	3	SEDANG	CEPAT
P	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	TELAT
L	LUAR KOTA	KEJURUAN	DALAM KOTA	2	SEDANG	CEPAT
L	LUAR KOTA	KEJURUAN	DALAM KOTA	2	SEDANG	CEPAT
L	DALAM KOTA	UMUM	DALAM KOTA	2	SEDANG	TELAT
L	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	TELAT
L	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	TELAT
P	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT

Step 2 is based on the training data classified student data given attribute gender, city of birth, type of school, city schools, 27 welfare economics, and the GPA method Bayres Naive Classifier. Data test: Gender: Male City of birth: City of Inner Type of school: Vocational State School: City of Inner GPA: 3 Economy: Moderate Step 3 Based on the test data, it can be decided with a few rare: Phase Counting the number of classes or label P (fast) = 21/50 number of training fast data shared by all the amount of data P (right) = 6/50 of training appropriate amount of data shared by all the amount of data P (late) = 23/50 training late the amount of data shared by all of the amount of data Stage count the number of the same case with the same class: P (Gender = Male | Y = Fast) = 15/21 P (Gender = Male | Y = right) = 4/6 P (Gender = Male | Y = late) = 20/23 P (City Inner City birth = | Y = Fast) = 5/21 P (City Inner City birth = | Y = right) = 3/6 P (City Inner City birth = | Y = late) = 5/23 P (mode = Vocational Schools | Y = Fast) = 5/21 P (mode = Vocational Schools | Y = right) = 2/6 P (mode = Vocational Schools | Y = late) = 23/7 28 P (City of Inner City School = | Y = Fast) = 11/21 P (City Inner City School = | Y = right) = 1/6 P (City Inner City School = | Y = late) = 8/23 P (GPA = 3 | Y = Fast) = 16/21 P (GPA = 3 | Y = right) = 5 / 6 P (GPA = 3 | Y = late) = 14/23 P (Economy = Medium | Y = Fast) = 19/21 P (Economy = Medium | Y = right) = 5/6 P (Economy = Medium | Y = late) = 23/23

Phase multiplied all variable results quickly, accurately and late. P (Men \ Quick) x P (Inner City \ Quick) xp (SMK \ Quick) x P (2.8 \ Quick) x P (Inner City \ Quick) x P (Medium \ Quick). = 15 21 5 21 5 21 11 21 16 21 19 21 = 0.7143 0.2381 0.2381 0.5238 0.7619 0.9048 = 0.0146 P (Men \ Right) x P (City Inner \ Right) x P (SMK \ Right) x P (2.8 \ Right) x P (Inner City \ Right) x P (Medium \ Right). = 4 6 3 6 2 6 1 6 5 6 5 6 = 0.0667 0500 0.3333 0.1667 0.8333 0.8383 = 0.0129 29 P (Men \ Late) x P (City Inner \ Late) x P (SMK \ Late) x P (Inner City \ Late) x P (2.8 \ Late) x P (Medium \ Late). = 20 23 5 23 7 23 8 23 14 23 23 = 0.8696 0.2174 0.3043 0.3478 0.6087 1 = 0.0122 Phase compare a result of the rapid, accurate, and late. From the results of our calculations, can be seen in the calculation was the highest score belongs to a class probability value (P | Quick), so that it can be concluded students can graduate quickly. Based on the results of the implementation using the 50 data training or training of any class of a percentage amount in the category 42% faster, 12% fall into the right category and 46% in the category of late, late category received a high percentage. Then the next step doing testing to the data 20 Data testers then obtained 20% faster category, 35% of the category, and 45% category of late. According to the journal is said that the test data, the result is that women tend to pass more quickly than the dominant male did not graduate on time or late. d. Journal of Personality Types Indonesian Classification for Text in Partners SearchingSAT Website Using Naïve Bayes Methods This journal has a case of almost all people do not know the personality of each so difficult to match her personality with her partner. Solutions for knowing personality ourselves and get a matching pair with personality are using the Naive Bayes classifier. Data used in this journal 30 of 40 the document data 160 training and learning data record. Classified into four categories, namely: sanguine, koleris, melankonis and plegmatis. The following discussion of the calculation process Naive Bayes classifier on this journal Step 1 phase document the learning phase, this process is used to obtain a probabilistic value of ( )and (|).Learning flowchart Figure 8

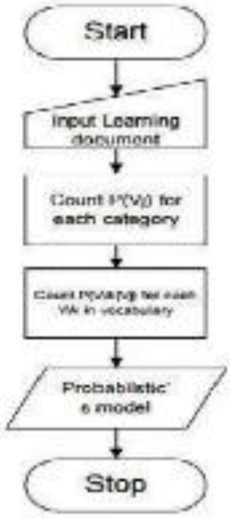

Start
Input Learning document
Count P(Vj) for each category
Count P(Wk|Vj) for each Wk in vocabulary
Probabilistic's model
Stop

Figure 7. Naive Bayes Classifier Learning Flowchat

Next, calculate ( )for each category using the formula: ( )| ( )| | | Description: ( ):the number of words in the category and | | : Is the number of documents used in training.Next, calculate) |(for each in vocabulary by the formula:) |(=) |(+ 1 + | | Description: (|): is the number of eventssaid in the category : is the number of all words in the category and | | : Is the number of unique words (different) from all of the training data. Step 2 Stage classification is the phase where the new document will undergo a classification process based on previous data. Flowchart for classification phase can be viewed pad Figure 4.

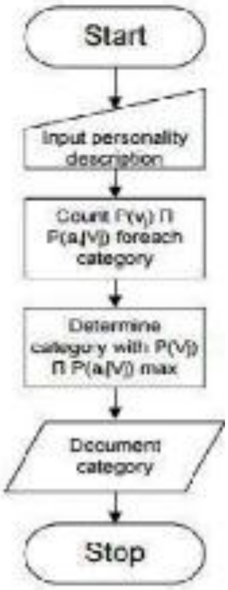
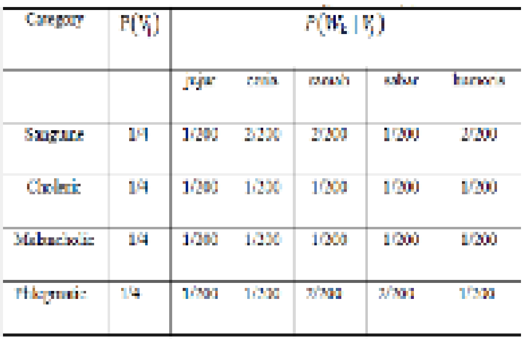

Start
Input personality description
Count P(vj) Π P(ai|Vj) foreach category
Determine category with P(Vj) Π P(ai|Vj) max
Document category
Stop

Figure 8. Process Naive Bayes classifier

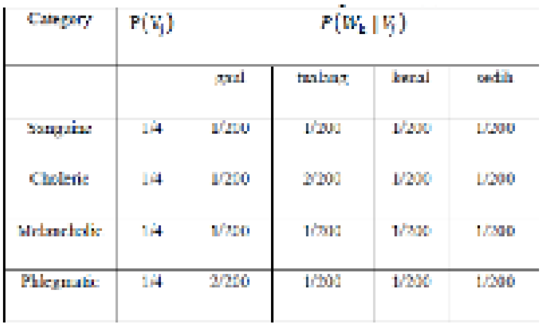
in the classification process, personality input document and probabilistic models that have been generated in the learning phase. Stage further with the formula: = ( )Π)|(Having obtained the calculation for each category, the category of the selected category maximumused to classify documents personality. The categories used in the classification of types of four personalities, namely: 1) Sanguin is a personality trait that clever persuasive and want to be famous, someone who has the personality is usually nicknamed the "popular" 2) Koleris is a personality trait that is often dominant and competitive, someone who have this personality usual nicknamed the "strong" 3) Melankonis a perfectionist personality traits and orderly, someone who has the personality is commonly nicknamed the "perfect". 4) plegmatis the personality traits of the faithful and avoid conflict, someone who has the personality is commonly nicknamed the "pacifist". This method depends on prior knowledge given. for example, the user inputs the data of personality, as follow: "I am an honest forest, cheerful, friendly, patient, and humorous. I like hanging out with friends. I like adventure. I love getting to know new things but I Also often sad". Results will calculate the Naïve Bayes method as shown in Table 14 and 13.

Table 14 Process 1Text Mining



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
Cholerik	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

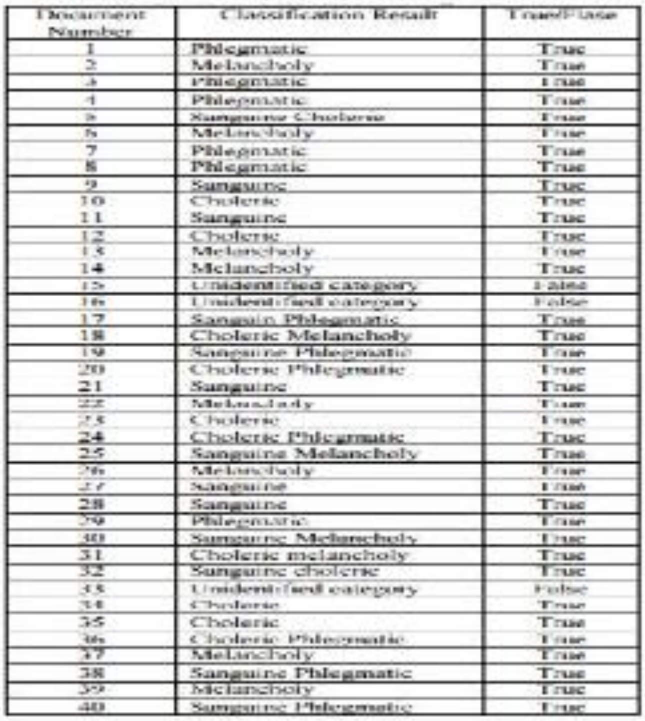
Table 15 Text Mining Process 2



Category	$P(V_j)$		$P(W_k \mid V_j)$		
		gaul	tenang	kesal	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 ( )and (|)then calculates category. P (optimistic | document) = 1 4 / 1200 / 2200 / 2200 / 1200 / 2200 / 1200 / 1200 / 1200 / 1200 / = 3.09 10-21 P (choleric | document) 34 = 1 4 / 1200 / 1200 / 1200 / 1200 / 1200 / 1200 / 2200 / 1200 / 1 200 / = 2 (2.048 1021) / = 0.997 10-21 P (melancholy | document) = 1 4 / 1200 / 1200 / 1200 / 1200 / 1200 / 1200 / 1200 / 1200 / 1200 / = 1 (2.048 1021) / = 0.488 10-21 P (apathy | document) = 1 4 / 1200 / 2200 / 2200 / 1 200 / 2200 / 1200 / 1200 / 1200 / 1200 / = 8 (2.048 1021) / = 0.09 10-21 text mining results with methods Naive Bayes classifier for documents calculated earlier is sanguine personality and phlegmatic. Once you know the type of personality we then find potential mates based on personality. As the result was known that a person has a personality mix, optimistic, and aptis. 35 As the theory of compatibility of couples, optimistic are a couple of melakonis and apathy are a couple of offense. So the partner should be a person who has melakonis, and irritability. This experiment uses the document data 40 160 training and learning data document, in Table 3, is the result of a detailed classification of personality for 40 training data that has been classified. There are 3 errors (error) which has three categories are not known. So that the percentage of errors is reached, the percentage accuracy: =100% (4) = 3740 100% = 92.5%

Table 16. Results of the classification of the training document



Document Number	Classification Result	True/False
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

3.2 Discussion

Based on the four journals have been described that contain cases and results calculations using Naive Bayes classifier method, the authors could analyze that this journal has four stages, the advantages and disadvantages of different. Stages, the advantages and disadvantages can be seen as follows:

A. Journal of Classification and Search Reference Book Academic Method Using Naive Bayes classifier (NBC) Stages methods Naive Bayes Classifier in this journal are as follows: 1) Phase Classification Text 2) Stage Text preprocessing 3) Stage Case Folding 4) Phase tokenizing 5) Phase Transformation text 6) Filtering Phase 7) Phase Stemming 8) Training process Stage 9) Results Pros: In the classification stage to the first stage up to five this journal provides clear explanations and easy to understand. Disadvantages: In the fifth step the journal is a journal does not display the calculation process test data that has been provided, which is displayed only on the calculation formula and description method of document classification stage 37 Naive Bayes classifier, so that these journals are not easily understood in its calculations.

B. Journal of Local News Classification Radar Malang Using Naive Bayes Method With N-Gram Features The steps of the Naive Bayes Classifier method in this journal are as follows: 1) Mining Text Stage 2) Pre processing Text Stage 3) Phase Case Folding 4) Tokenizing Phase 5) Phase Filtering 6) Stemming Stage 7) Stage of N-Gram Modeling 8) Testing Process Phase 9) Excess Results: In this journal the test is done several times the test with the amount of test data and training data are different, so can know the accuracy of the amount of data needed in the use of the Naive Bayes Classifier method in solving cases in this journal. Disadvantages: In this journal does not show the calculation process using Naive Bayes Classifier method but by displaying the final result only, thus making the reader difficulty in knowing the calculation steps Naive Bayes method.

C. Journal of Implementation of Naive Bayes Classification Method For IBI Graduation Time Prediction Dharmajaya Scholar The steps of Naive Bayes Classfier method in this journal are as follows: 1) Text Classification Phase 2) Calculation phase of class number 3) Phase calculation of the same number of cases with the same class 4) . The results of the description, with details and display the results of calculations, so that the journal is clear in the sense of calculation and easy to understand. Disadvantages: In this journal there is no pre processing text stage and the transformation text stage, which is just a classified stage directly used into test data.

4. Conclusions and Suggestions

4.1 Conclusions

From the results of the authors do to this research the authors can draw some conclusions related to the research process and with the content of the research itself, the conclusions obtained in this writing are: a. Of the four journals that the authors carefully, it is concluded that the journal describing the steps of the method of calculating the method of Naive Bayes Classifier with clear, easy to understand and accurate results is the journal Implementation Of Naive Bayes Classification Method To Predict Graduation Time Of IBI Dharmajaya Scholar. b. By applying Data Mining using the Naive Bayes Classifier (NBC) method, this method can help predict the future based on past or past events, so it can easily find out the future events of the graduates; can classify the news that needs to be submitted by the public and the unnecessary news delivered by the community because not all news has its accuracy and authenticity.

4.2 Suggestions

Suggestions in this study is the next research is expected the author get more data and information because the more data used then the results will be more accurate and relevant, becomes a valuable and useful information for the future, to get new information and useful in the future, need build an application system by applying the Naive Bayes Classifier (NBC) method to make it easier for people, organizations or companies to find useful data and information from a lot of data, quickly finding new information without having to do it manually.

References

[1] Agus Setiawan, et al, 2013. *Classification and Searching of Academic Reference Books Using the Naive Bayes Classifier Method (NBC)*. Journal of Informatics Mulawarman, **Volume 10** No. 1, 1-10.

[2] Denny Nathaniel Chandra, et al, 2016. *Classification Local News Radar Malang Using Naive Bayes Method With N-Gram Feature.* Journal of Scientific Technology and Informasia ASIA (JITIKA). **Volume 10** No. 1, 1-9.

[3] Ketut Artaye, 2013. *Implementation Of Naive Bayes Classification Method To Predict Graduation Time Of IBI Dharmajaya Scholar*. International Conference on Information Technology and Business. ISSN 2460-7223. 1-4.

[4] Ni Made Ari Lestari, 2013. *Personality Type Classification for Indonesian Text in Partners Searching Using Naive Bayes Method Website*. International Journal of Computer Science Issues (IJCSI). ISSN 1694-0784, 1-8.

[5] Dennis Aprillah C, et al, 2013. *Learn Data Mining with Rapid Maner*. Jakarta. Page 42-43.

[6] Jeffrey Stanton, 2012. *Data Science*. *Syracuse University*. Page 9.

[7] Tom M Mitchell, 2013. *Generative and Discriminative Classifiers Naive Bayes and Logistic Regression*. page 2.

[8] Rika Rosnellly, 2012. *Concepts and Expert System Theory*. Page 79.