Content Abstract Classification Using Naive Bayes

Syukriyanto Latif, Untung Suwardoyo, Edwin Aldrin Wihelmus Sanadi

Electrical Engineering Department, Hasanuddin University, Makassar, Indonesia
E-mail: syukriyanto24@gmail.com

Abstract. This study aims to classify abstract content based on the use of the highest number of words in an abstract content of the English language journals. This research uses a system of text mining technology that extracts text data to search information from a set of documents. Abstract content of 120 data downloaded at www.computer.org. Data grouping consists of three categories: DM (Data Mining), ITS (Intelligent Transport System) and MM (Multimedia). Systems built using naive bayes algorithms to classify abstract journals and feature selection processes using term weighting to give weight to each word. Dimensional reduction techniques to reduce the dimensions of word counts rarely appear in each document based on dimensional reduction test parameters of 10% -90% of 5,344 words. The performance of the classification system is tested by using the Confusion Matrix based on comparative test data and test data. The results showed that the best classification results were obtained during the 75% training data test and 25% test data from the total data. Accuracy rates for categories of DM, ITS and MM were 100%, 100%, 86%, respectively with dimension reduction parameters of 30% and the value of learning rate between 0.1-0.5.

1. Introduction.

Technological developments have resulted in the increase of large amounts of data. Large data that has been stored is rarely used optimally because of humans often do not have sufficient time and ability to manage them. Data large volumes like text data, goes far beyond the ultimate human processing capacity limited.

The case in detail is the classification of abstract content in English. There are 3 areas of abstract content categories used as training data and test data, namely Intelligent Transport System, Data Mining and Multimedia. But abstract content labeling is currently still done manually, thus taking a long time in knowing which category of abstract content is. To solve it. Automatic content category system needs to be required, to minimize errors. Data mining approach has been widely used in many aspects of especially evolving data series[1-3]

As for some research on classification of documents that become reference source in this research is classification text with naive bayes classifier to classify news text and abstract academic by amir hamzah in year 2012, with method naive bayes can help user in classification with acquisition rate of 91% for text news and 85% for academic abstracts [4].

Further research conducted by Ashis Kumar Mandal et al in 2014, apply 4 algorithms DT (Decision Tree), KNN (K-Nearest Neighbors), NBC (Naive Bayes Classifier) and SVM (Support Vector Machine) in classifying web documents. The accuracy is 80.65%, KNN 74,24%, NB 85,22% and SVM 89,14%, research of G. Keerthika and D. Saravana Priya in 2015 apply NBC method by combining 2 feature selection said GR (Gain Ratio) and IG (information Gain) combined these 2 methods are able to produce accuracy of 92% [5].

In G. Keerthika and D. Saravana Priya in 2015, applying the NBC method by combining 2 features of the said GR (Gain Ratio) and IG (information Gain), combined these two methods are capable of achieving an accuracy of 92% [6]. Amalia Anjani, Djumardi et al in 2015 also apply NBC Algorithm (Naive Bayes Classifier) with Confix Stripping Stemmer as the classification of Indonesian news article. Based on the validation results using 10 Cross validation NBC achieved the best accuracy of 86.97% [7]. Fajar Rohman Hariri et al in 2015 also conducted a thesis abstract classification research using the method of learning vector quantization by dividing 3 categories of fields namely SI RPL (Software Engineering Information System), CAI (Computation - Artificial Intelligence) and Multimedia, in...
feature selection research used namely dimensional reduction so as to increase accuracy in the validation process. 100% accuracy obtained in SI RPL and CAI, 70% for Multimedia interest area, so that overall accuracy can be 90%, best condition get with dimension reduction parameter 20% [8]. I Gusti A Socrates et al in 2016 conducted a study using the NBC method by applying the Gain Ratio weighting feature selection in classifying Indonesian text, the accuracy of NBC reached 91% [9].

From some of the above research, the authors propose to use naive bayes method and Term Weighting algorithm also dimension reduction on the feature selection so that the process of classification accuracy gets better.

2. Methodology
This research is an experimental research because to get the performance best algorithm naive Bayes in classifying, conducted several times experiments with different parameters. The abstract content data can be on the www.computer.org page. From 3 categories of Intelligent Transport System, Data mining and Multimedia. There are 120 abstract content file data with pdf extension. For data analysis, from the abstract document obtained, will be experimented with several trial scenarios described in Table 1:

Table 1. Scenario test.

| No | Amount of data train | Amount of data testing | Information |
|----|----------------------|------------------------|-------------|
| 1  | 100%                 | 100%                   | The entire abstract content becomes data learning as well as data testing |
| 2  | 75%                  | 25%                    | 75% of the data will be used as data learning, and the remaining 25% is used as data testing. |
| 3  | 50%                  | 50%                    | 50% of the data will be used as data learning, and the remaining 50% is used as data testing. |
| 4  | 25%                  | 75%                    | 25% of the data will be used as data learning, and the remaining 75% is used as data testing. |

It will also be analyzed by different dimension reduction with dimension reduction process, the amount of data learning and data testing that are different in expect to know the best accuracy result from Naive Bayes method in classifying abstract content. The design of the proposed system is shown in Figure 1.

Figure 1. Proposed system design.
2.1 Preprocessing
There are several steps that need to be done in this process is case folding that change the letters in documents into lowercase, tokenizing as input string cutter based on each word that make up, filtering take important words but in this research use stoplist algorithm (throw less word important) and stemming is the stage of searching for root words (base word) of each word.

2.2 Feature Selection
In this stage there are 2 processes that will be done that is Term Weighting is giving the term weight that appears on each abstract content and dimension reduction by removing words that will only appear in one or a few on the abstract content. The following shows the implementation of the results of the Term Weighting word in Figure 2.

| Document Training | Word And Term Frequency |
|-------------------|-------------------------|
| A collision prevention mechanism for the multicast transport in IEEE 802.11 networks | ag => 1, based => 1, bs => 2, busy => 1, caused => 2, collision => 4, compliant => 1, content => 1, define => 1, deliver => 1, deployed => 1, exceed => 1, factor => 1, failure => 1, fair => 1, now => 1, guarantee => 1, ieee => 2, interfere => 1, introduce => 1, loss => 2, lost => 1, measure => 1, mechanism => 2, medium => 1, missing => 1, mode => 1, modification => 1, multicast => 4, network => 2, packet => 4, paper => 1, path => 1, perfective => 1, policy => 1, prevent => 1, rate => 1, receiver => 1, require => 1, retransmitted => 1, retry => 1, sharing => 1, show => 1, standard => 2, study => 1, symbol => 1, time => 1, transmission => 1, transmitted => 1, transport => 2, unicast => 1, wlan => 1 |
| A composition toolkit for experimental wireless transport processes | algorithm => 1, architecture => 1, area => 1, compile => 1, component => 1, conduct => 2, control => 1, decompose => 1, demonstrate => 1, describe => 1, develop => 2, developing => 1, differ => 1, easier => 1, effort => 1, enable => 1, evaluate => 1, expert => 2, experimental => 2, favorably => 1, greatly => 1, illustrate => 1, implementate => 1, improving => 1, kernel => 1, linux => 1, multiloop => 1, network => 2, opensource => 1, paper => 1, perform => 3, performed => 1, plugndisplay => 1, present => 1, previously => 1, protocol => 4, rapid => 1, race => 1, releasing => 1, reliable => 1, reported => 1, required => 1, research => 2, result => 1, show => 1, significantly => 1, simplifying => 1, smaller => 1, tcp => 1, toolkit => 2, transport => 2, wireless => 4, user.space => 1, utility => 1, variance => 1, wifi => 5, wireless => 4, work => 1 |
| A flash transportation layer for multimedia storages | accessed => 1, address => 1, application => 1, based => 1, block => 1, called => 1, collection => 1, data => 1, device => 3, offer => 1, drive => 1, experimental => 1, fti => 3, file => 1, filter => 1, filtering => 1, flash => 2, frequently => 2, fti => 2, garbage => 1, idea => 1, improve => 1, index => 1, layer => 1, main => 1, manage => 2, manner => 1, mapping => 1, metadata => 2, multimedia => 2, hand => 1, overhead => 1, pagelevel => 1, paper => 1, perform => 1, propose => 1, proposed => 1, read => 1, reducing => 1, request => 2, result => 1, scheme => 5, separately => 1, sequential => 1, server => 1, show => 1, significantly => 1, solidstate => 1, ssd => 1, storage => 3, term => 1, tradition => 1, transition => 1, update => 1, updated => 2, user => 2, widely => 2 |

Figure 2. Term Weighting.

Dimensional reduction process is done to minimize the dimension of data so that computation time is needed less. However dimensional reduction processes must pay attention to data characteristics, since the missing dimensions may also eliminate the data characteristics [10]. Therefore in this study selected dimensional reduction by removing sparse part. Term that has a frequency that is mostly zero will be eliminated, since the term represents a very rare term in the data or in other words term may be ignored.

The amount of dimension dimension of the term in accordance with the parameters specified. Because it is feared that this process will eliminate the characteristics of data, then created value parameters that serve to limit dimensional reduction. So even if in a term most values are zero, the term will not be omitted if any of its frequency values exceed the specified parameter.

TF data that has been entered in the database has different values. The term that has a zero value TF is negligible, with no TF values exceeding the specified parameters. In other words dimensional reduction process is by eliminating the feature or term that the frequency of all / mostly worth 0. If there are some that are not worth 0 then the TF value must exceed certain parameters. This is so that
data does not lose its characteristics. So the process can run well. The following dimensional reduction illustration is shown in Figure 3.

![Dimensional Reduction Illustration](image)

**Figure 3.** Illustration of dimensional reduction.

In Figure 3, the parameter is shown as 40%. Then dimensional reduction is done by omitting a column whose value is not 0 less than 40% of the data. The 5-9 columns will be removed because the column is not eligible.

Of the 120 abstract content data after being processed it produces 5,344 terms (words). Of the 5,344 words available, there are some words that appear only in one or two abstract content only, so the word can be ignored. Therefore, to eliminate these words, and also to speed up the calculation process, a dimensional reduction process is performed. The result of the number of terms produced after the dimensional reduction process is shown in Table 2 below:

| Dimension Reduction Value | Number of Term |
|---------------------------|----------------|
| 10%                       | 95             |
| 20%                       | 70             |
| 30%                       | 39             |
| 40%                       | 25             |
| 50%                       | 14             |
| 60%                       | 8              |
| 70%                       | 3              |
| 80%                       | 3              |
| 90%                       | 1              |

The % value above means the words taken are words that appear in more than how many% of documents. The fewer parameters, the less the term will be generated.

### 2.3 Classification

From the weighting result of Term Weighting and dimension reduction will then be used as training data and test data on Naive bayes classifier process. The naive bayes process includes two stages: learning stages and testing stages. The main characteristic of the naive bayes classifier is a very
strong (naive) assumption of the independence of each condition or event. The Naive Bayes Classifier is a simplified model of a suitable bayes algorithm in the classification of texts or documents as in the following equations:

\[
P(C_i \mid W_k) = \frac{P(W_k \mid C_i) \times P(C_i)}{P(W_k)}
\]

Where,
- \(P(C_i \mid W_k)\) is the probability of occurrence of categories \(C_i\) with the word \(W_k\)
- \(P(W_k)\) "Constant" for all categories so it is only formed \(P(W_k \mid C_i) \times P(C_i)\) which needs to be maximized
- \(C_i\) is the category available (\(C_1, C_2, ..., C_l\))
- \(P(C_i)\) is the probability of occurrence of categories \(C_i\)
- \(P(W_k \mid C_i)\) is the probability of occurrence of words \(W_k\) in categories \(C_i\)

So the equation for solving this problem is as follows:

\[
P(\text{words} \mid \text{category}) \frac{P(\text{words} \mid \text{category}) P(\text{category})}{P(\text{category})}
\]

a. Training process
In the training phase, the process of analyzing the sample of documents in the form of vocabulary selection, which is a word that may appear in the collection of sample documents that can represent the documents as much as possible, as well as the formation of classes and as a reference how abstract content will be classified. to model the probability of the training process the following equations are used.

\[
P(c_i) = \frac{\sum_{D} f_d(c_i)}{n}\]

\[
P(w_k \mid c_i) = \frac{n_{k+1}}{n + |\text{vocabulary}|}
\]

where.
- \(P(c_i)\) is the amount of abstract content that has categories \(c_i\)
- \(D\) is the sum of all abstract content training.
- \(n_k\) the value of word occurrence in each category \(c_i\)
- \(n\) is the total number of words in a category \(c_i\)
- \(|\text{vocabulary}|\) is the total number of words.

So the equation for modeling the probability of the training process is used by the following equation.

\[
P(\text{category}) \frac{\text{Abstract Content Amount in category}}{\text{The sum of all data is training abstract content}}
\]

\[
P(\text{words} \mid \text{category}) \frac{\text{The value of word occurrences in each category} + 1}{\text{Number of all words in category} + \text{number of all data coaching words}}
\]
b. The testing process
   This stage as the core of the naive bayes is to know the accuracy of the model built on the training process to predict the unknown class label
   • Generate probabilities for each class according to equation (1) using the P (category) and P (category | categories) that have been obtained from the training.
   • The maximum class probability value is the category of the selected class as the result of classification.

3. Results
After dimensional reduction process, the training process is done using Naive Bayes Classifier. From the final weights generated, then predicted by calculating the probability of the occurrence of the word testing on the NBC method. With 100% data used as data testing as well as training data and with different dimension reduction scenarios, it produces accuracy as in Table 3 below:

| Dimension Reduction Value | Accuracy (%) |
|---------------------------|--------------|
| 10%                       | 68.56        |
| 20%                       | 76.66        |
| 30%                       | 88.48        |
| 40%                       | 65.33        |
| 50%                       | 58.45        |
| 60%                       | 51.88        |
| 70%                       | 51.88        |
| 80%                       | 51.88        |
| 90%                       | 51.88        |

From the above tables and graphs, the dimensional reduction of 60% -90% yields only 51.88% accuracy, all the abstract data are classified into the field of multimedia interest, and from table 3 it is known that the best accuracy is obtained when 30% dimension reduction with accuracy reached 88.48%. With the introduction details for each category are listed in table 4:

| Category | The amount of content | The number of successful classified | accuracy  |
|----------|-----------------------|-------------------------------------|-----------|
| ITS      | 40                    | 38                                  | 96%       |
| DM       | 40                    | 34                                  | 88%       |
| MM       | 40                    | 30                                  | 80%       |

From the above table it is known that the best NBC method in classifying abstracts for ITS category with accuracy of confusion matrix is 96%, 88% in DM category, and lowest accuracy in classifying abstract for Multimedia category which only succeeds in classifying 30 abstract content from 40 content abstract and produce an accuracy value of 80%.
After that, trial scenario is done by changing the data of learning and testing. Accuracy is generated as in Table 5.
Table 5. Results Accuracy Scenario Total Data Training.

| Learning Rate | 75% Training | 50% Training | 25% Training |
|---------------|--------------|--------------|--------------|
|               | 5% Training  | 25% testing  | 75% testing  |
| 0.1           | 92%          | 86%          | 76,77%       |
| 0.2           | 92%          | 86%          | 75,55%       |
| 0.3           | 92%          | 86%          | 76,89%       |
| 0.4           | 92%          | 86%          | 76,56%       |
| 0.5           | 92%          | 86%          | 76,56%       |
| 0.6           | 72,33%       | 81,33%       | 75,33%       |
| 0.7           | 60%          | 81,33%       | 74,33%       |
| 0.8           | 55,53%       | 65,66%       | 71,66%       |
| 0.9           | 65,50%       | 44,66%       | 61,66%       |

Table 6. Details of 75-25 accuracy and 0.1-0.5 learning rate.

| Category | The amount of content | The number of successful classified | Accuracy |
|----------|-----------------------|-------------------------------------|----------|
| ITS      | 40                    | 40                                  | 100%     |
| DM       | 40                    | 40                                  | 100%     |
| MM       | 40                    | 33                                  | 86%      |

From the above table it is known that the NBC method is good in classifying abstract for ITS and DM category with 100% accuracy, and abstract classification for MM category only yields an accuracy value of 86%. From several test scenarios with dimensional dimension value variation, the amount of learning and test data produces different accuracy. In the scenario of the amount of data learning and testing, the more data learning the better the accuracy. The best results were obtained by dimensional reduction of 30%, using 75% data learning, and from 25% of test data, the NBC method successfully classified 95.33% of the abstract content correctly. With the best result using learning rate between 0.1 – 0.5, can classify 100% in category of ITS and DM, but MM abstract content is only 86%.
4. Conclusion

From the result of the research, it is concluded that the accuracy produced by NBC method in classifying abstract content reaches 95.33% for all data, by successfully classifying abstract content in ITS category, DM with 100% accuracy and MM with 86% accuracy. The best accuracy is found in dimension reduction condition with 30% parameter, data testing and training ratio is 75% data testing, 25% training data and learning rate between 0.1-0.5.

References

[1] Indrabayu, Harun N, Pallu M S and Achmad A 2013 *International Journal Of Engineering And Technology*

[2] Sespajayadi A, Indrabayu, Nurtanio I 2015 *International Seminar on Intelligent Technology and Its Application,. ISITIA Proceeding*

[3] Sidehabi S W, Indrabayu ,Tandungan S 2017 *Proceedings- CYBERNETICSCOM: International Conference on Computational Intelligenceand Cybernetics*

[4] Hamzah A 2012 *National Seminar on Science & Technology Applications*) Periode III (Yogyakarta) ISSN:1979-911X

[5] Manda A K and Sen R 2014 *International Journal of Artificial Intelligence & Applications 5(5)*

[6] Keerthika G and Saravana P D 2015 *Journal of Network Communications and Emerging Technologies 1(1)*

[7] Amalia A A, Djunaidy A, Kusumawardani R P 2015 *Classification Articles News On The Net Based Naive Bayes Classifier Using Confix-Stripping Stemmer,( ITS-Undergraduate)-38449-5210100106*

[8] Fajar H R, Utami E and Amborowati A 2015*Citec Journal 2(2) ISSN: 2354-5771*

[9] Guan I, Socrates A, Afrizal L, Akbar, and Akbar M S 2016 *J.Lontar komputer 7(1)*

[10] Yang Y, Pedersen J O 1997 *Proceedings of the Fourteenth International Conference on Machine Learning* (Nashville, Tennessee, USA)