AUTOMATIC NEWS ARTICLES CLASSIFICATION IN INDONESIAN LANGUAGE BY USING NAIVE BAYES CLASSIFIER METHOD

Arni Darliani Asy'arie
Program Studi Teknik Informatika, Fakultas Sains dan Teknologi
Universitas Islam Negeri Syarif Hidayatullah Jakarta
Jl. Ir. H. Juanda No. 95, Ciputat 15412 Jakarta
0217493547
Email : darliani_dd@yahoo.com

Adi Wahyu Pribadi
Jurusan Teknik Informatika, Fakultas Teknik
Universitas Pancasila
Jl. Srengseng Sawah, Jagakarsa Jakarta 12640
0217864730
Email : adi.wahyu.p@gmail.com

ABSTRACT

Curently, internet content growth rapidly. Automatic news classification is the classification of news into a category. In this research, the classification method used is Naive Bayes wich is known as Naive Bayes Classifier (NBC). The classification process covers: case folding, parsing, stopword elimination, stemming, words weighting, and documents classification by using NBC. This research uses 250 news articles divided into 5 categories as learning documents. The trial results of this research shows that the system is able to generate such accuracy in delivering news articles classification with the average Recall value of 92.87% and Precission value of 91.16%.

Key word : Article, learning documents men pembelajaran, document classification, Naïve Bayes Classifier, stemming

1. INTRODUCTION

The need for information, such as in Indonesia has been growing along with time. One of the very valuable information source today is the electronic news media which able to get information quickly amidst someone's packed daily activities. One of the electronic media as stated above is the internet. Frequently updated news site characters can cause information to flow in huge amount every day. Automatic news classification, the classification of news into a category, is very much needed to analyze news in an effective and efficient way. One of the classification methods that we can use is the Navie Bayes or is commonly callled as Naive Bayes Classifier (NBC).

NBC is a classification method has high accuracy with simple calculation [4]. NBC has been the core of information rediscovery framework for years. NBC has proven its great performance, even when it is not at its optimum level its performance is still more optimum than other representational classifiers [6] such as K-means, Clustering and Cosine Similarity. The position of words in this method plays an important role in classification process, thus stemming method is needed (affixes elimination). Stemming is used to find the basic form of words with affixes [8]. Other factors affecting the words position is learning document because the method that will be used in the application construction is a part of supervised document, a classification method using the training phase. The purpose of this research is to build a software capable of determining categories of articles in Indonesian language.

There are several other news documents classification research using Naive Bayes Classifier classification, for examples are research conducted by Yudi Wibisono [13] and Sylvia Susanto [12]. Yudi Wibisono conducted a research on 582 news documents from the site of www.kompas.com without applying the stemming process. Sylvia Susanto conducted a research using news source from the site of www.suarapembaruan.com and the weighting method used is by counting the words appearance frequency.

2. DOCUMENTS CLASSIFICATION

Classification is a process in search of a batch of function documents that describe or differentiate data classes so that the process can be used to predict classes of a document which class was unknowned perviously. There are two main variants in document classifications, namely supervised classification document and unsupervised classification document. Supervised classification document is a document classification that has a learning method with training document in the form of learning document. While unsupervised classification document is a method applied independently without training nor teaching used to analyze structure and inter data relationships [10].

The construction of this application has a concept that covers two phases, namely learning phase and classification phase. Learning phase must be executed first to develop a learning document as a guideline for each category. Learning phase has a module that resembles the classification phase. The only difference is that learning phase does not execute a classification module, but only generate documents consisting of words to characterize a category. Classification phase can be specified through several process as stated in the following:

a. Input acceptance in the form of news document texts which category is not known before.
b. Afterwards is the process of document reprocessing, initiated by reading sentences from document texts. Read sentences will then go through a case folding. process that eliminates all characters execpt letters and spaces and all letters are converted into small leters. Next is the parsing process that makes sentences into independent words. Words fitting into stopwords list must be eliminated. After that stemming process with the aid of basic form of words dictionary in Indonesian language is conducted.
c. Words, after going through preprocessing documents phases, are then ranked with weighting process.
d. Furthermore, classification process conducted a calculation based on the Naive Bayes Classifier.

2.1. Preprocessing

The purpose of *document preprocessing* is to unify words, eliminate noise, and lessen the vocabulary volume. The steps are:

- Case folding is the conversion process of all letters in documents into small letters. Only letters “a” to “z” that can be accepted. Characters besides letters are eliminated and considered as delimiter [5].
- Parsing is a process to break down sentences into words. Elimination of stopwords eliminates frequently found terms that are not useful in information rediscovery, namely conjunctions.
- Stopwords elimination is the process of eliminating terms that frequently appears and are not useful in information rediscovery with the purpose of filter words that are found frequently and have low value and are not useful in information retrieval. Because usually more than 80% of the words in documents collection are useless words in information retrieval.

2.2. *Stemming*

process is a process of cutting or eliminating affixes in a word. The affixes variants are prefixes, suffixes, infixes, and confixes (the combination of prefixes and suffixes). Stem is the part of the words after affixes have been eliminated.

Stemming rules used are as follows:

1. ParS (eliminate –kah, -lah particles)
2. ProS (eliminate substitute words of, -ku, -mu, -nya)
3. PreS1(eliminate se- suffix)
4. ConS (eliminate ke-an, ke-)
5. SufS2 (eliminate –kan suffix)
6. SufS3 (eliminate –an suffix)
7. PreS2 (eliminate mem-, me-, meng-, meny-, men-, and di-prefixes)
8. ConS (eliminate ke-an, ke-i)
9. SufS1 (eliminate –man, -wan, -wati suffixes)
10. PreS3 (eliminate ber-, be-, bel- prefixes)
11. PreS4 (eliminate pem-, peng-, peny-, pen-, per-, pel-, ter-, and te-prefixes)
12. FSuf1 (eliminate –sionis, -isme, -itas, -isasi suffixes)
13. FSuf2 (eliminate –if, -ik, -is suffixes)
14. FSuf3 (eliminate -at, -wi, -al, -wiah, -iah suffixes)
15. FSuf4 (changes absorbed words as in words with ended with a vowel letter + -v, -v are changed to –f and in words ended with a vowel letter + -pt, -kt, -nt, eliminate –t)
16. SufS4 (eliminate suffix –i if the previous letter is not –i and –ng and or double consonants.

2.3. Words Weighting

In this phase, every words that have been processed through preprocessing document phase then its weightings is computed to generate words that can represent a category. Weighting calculation on corpus documents uses the formula of:

TF = 1 +ln (tf).

And words weightings of input documents uses the formula of:

(tf t, d) = appearance frequency of the term t in document d.

2.4. *Naive Bayes Classifier* Classification

In NBC, every news document is represented in attributive pairs (a_1, $a2_2$ .... a_n) where a_1 is the first word, a_2 is the second word, and so on. While V is news category compilation (sports, technology and some others). During classification, the Bayes approach will generate category label with highest probability (argmax) with attributive input (a_1, a_2 .... a_n). The following is the NBC calculation formula

:

$$v_{MAP} = \arg\max_{V_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

P(V_j) [illegible] ry of P(a_i | V_i) is determined by using the equation of:

$$P(V_j) = \frac{|\text{Category}_j|}{|\text{Corpus}|}$$

$$P(a_i \mid V_j) = \frac{1 + n_i}{n + |\text{Vucabulary}|}$$

- | category $_j$ | : the number of words ot input document i in category j.
- | corpus | : the number of documents in a learning document.
- n_i : is the appearance of word *ai* in category *vj*.
- | vocabulary | : the number of unique words in all learning document.

2.5. Implementation

Classification building process in this research is divided into three sub-processes as follows:

a. Document Filter

This process is conducted in learning phase and classification phase. This process is a process of eliminating noises in documents which is done by unifying words into basic words with small letters and only several important words (excluding conjunctions).

b. Calculating Word Weightings

This process is carried out by calculating word weightings on each document. The word weighting calculation process in learning phase is different than in classification phase:

1. In learning phase, the weighting calculation uses the TF Logarithm formula and the result will be stored into the data basis in the system.
2. In classification phase, the weighting calculation is carried out by counting the word frequency numbers in the input document.
   From the generated weighting results, the word taken is the word on the top 25% of the highest value.

c. Calculating Classification Value

In this phase, Naive Bayes Classifier value is conducted on the input document to generate category of an article automatically.

Learning Phase

Learning document contains 250 articles which are divided into five categories executed on processes in the learning phase and that generates words having characteristics in each category. Below is the result generated from learning phase:

Category	Total of Documents	Total of Terms	Jumlah kata dari 25% nilai TF tertinggi
Economy	50	8388	520
Sport	50	6050	491
Teknologi	50	8766	633
Health	50	10593	634
Metropolitan	50	5515	435

Table 3.1. Result of Learning Phase

The 25% highest in number of words in all categories are = 2713.

Classification Phase

Classification phase is a phase conducted to generate end result in the form of a category of a news article. After all documents have gone through document preprocessing and weighting process, accordingly reference words are generated and they will be processed using the Naïve Bayes Classifier on the input document to obtain a category of an article automatically

3. TRIAL AND EVALUATION

Documents collection used as learning document are taken from online news articles in the KOMPAS daily. The total learning documents used are 250 documents in five categories, that is economy, sports, health, technology, and health

3.1. Trial

Trial process on automatic article classification application is conducted by running the process using several trial documents ranging from 12-60 documents in Indonesian language from the five categories indicated in the above. Trial documents are obtained from three online news media, such as www.kompas.com, www.republika.com, and www.suarapembaruan.com and 20 news documents are taken from each site.

3.2. Evaluation

Evaluation on documents classification result on each category are conducted by using the standards in table 2.1 containing various possibilites of classification results on each category.

Category C_j		*Expert Judgments*	
		YES	NO
Classifier Judgements	YES	*a*	*d*
	NO	*b*	*c*

Table 32. Category of Classification's result

Table 3.2 indicates that classification results are sometimes inline with the expert decisions (*a*) and sometimes not (*b*). While documents not included in one of the categories are usually can not be classified into that category (*d*), and sometimes should be included in that particular category (*c*).

The four parameters in the above tables used to calculate 3 evaluation methods are:

a. *Recall* which is a comparison between relevant documents can be known with the total of relevant documents [3]. *Recall* has a formula that can be stated in below equation

$$R = \frac{a}{(a + c)}$$

b. *Precision* is comparison between relevant documents known with several known documents [3]. *Precision* has a formula that can be seen in equation

$$P = \frac{a}{(a + b)}$$

c. . *F-measure* is a value representing the whole system performance which is the combination of *recall* and *precision* value. *F-measure* has a formula that can be seen in the following equation

$$\text{F-measure} = \frac{2PR}{P + R}$$

Accuracy value from the Naive Bayes Classifier method can be divided into 5 document proportions with 12 documents multiple along with the documents remarks as follows:

Document	Total of Cocuments	Relevant Document
Economy	12	10
Sport	12	11
Technology	12	9
Health	12	11
Metropolitan	12	7

Table 3.3. Trial Document

With calculation :

- 12 proportion documents (economy) →
  Recall = 10/(10+1) * 100%
  = 90%
  Precison = 10/(10+1) * 100%
  = 90%
  F-measure = 2 * 90% * 90% / (90% + 90%)
  = 90%

- 24 proportion documents (economy + sport) →
  Recall = 21/(21+1) * 100%
  = 93,4%
  Precison = 21/(21+23) * 100%
  = 91,3%
  F-measure = 2 * 93,4% * 91,3% / (93,4% + 91,3%)
  = 92,33%

- 36 proportion documents (economy + sport + technology) →
  Recall = 30/(30+2) * 100%
  = 93,75%
  Precison = 30/(30+3) * 100%
  = 90,9%
  F-measure = 2 * 93,75% * 90% / (93,75% + 90,9%)
  = 92,33%

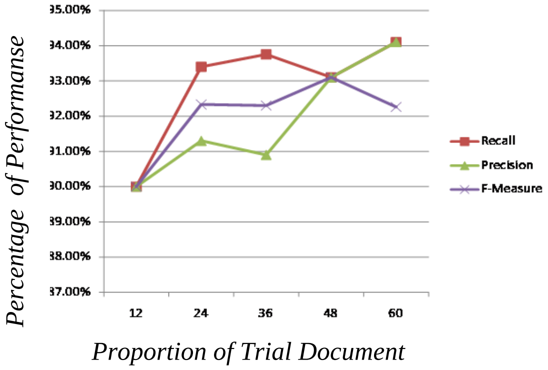
- 48 proportion documents (economy + sport + technology + health) →
  Recall = 41/(41+3) * 100%
  = 93,1%
  Precison = 41/(41+3) * 100%
  = 93,1%
  F-measure = 2 * 93,1%* 93,1% / (93,1%+ 93,1%)
  = 93,1%

- 60 proportion documents (economy + sport + technology +health + metropolitan) →
  Recall = 48/(48+3) * 100%
  = 94,11%
  Precison = 48/(48+5) * 100%
  = 90,5%
  F-measure = 2 * 94,11%* 90,5% / 94,11%+ 90,5%)
  = 92,26%

Proporsi Document	Recall	Precision	F-measure
12	90%	90%	90%
24	93,4%	91,3%	92,33%
36	93,75%	90,9%	92,30%
48	93,1%	93,1%	93,1%
60	94,11%	90,5%	92,26%

Table3.3. Performance of NBC by some proportion documents

Below is the description of the performance of Naïve Bayes Classifier::



3.1. Performance of NBC

From the results in the above, it can be seen that the Naive Bayes Classifier accuracy value is relatively high on lowest document proportion and also on highest document proportion, specifically all above 90%.

4. CONCLUSION

1. The process needed in conducting indexing on a document is by conducting the Case folding, parsing, stopword elimination, and stemming process.
2. The method used in calculating indexed word rank levels is by using the TF method. On the learning phase, the TF method used is the logarithm method while in classification phase, the method being used is a pure TF.

3. According to the trial result, it can be concluded that NBC method is a classification method with high accuracy level. This can be proven from the comparison of the system result with all the categories from several online news sites with all average values above 90%.
4. In this research, the input documents that can be processed by the system are documents with vocabulary not exceeding than 1,000 words. Because the large number of words will only give a null value as a result. For further research, it is hoped that the system can process documents with large volume of words.
5. There are several weaknesses in this research, that is each input document will certainly be classified into one of the available categories, though that particular document actually should not be classified into one of the categories in the system. It is hoped that the next research will be able to address this matter thoroughly.

5. REFERENCES

[1] Arifin, Zainal Agus dan Setiono, Novan Ari . *Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering.* Fakultas Teknologi Informasi Institut Sepuluh November (ITS) Surabaya. 2005. http://mail.its-sby.edu/~agusza/SITIAKlasifikasiEvent.pdf diakses tanggal 4 Februari 2008

[2] Baeza-Yates, Ricardo dan Berthier, Ribeiro-Neto . *Modern Information Retrieval.* Addison-Wesley. New York. 1999

[3] Grossman, David A dan Frieder, Ophir. *Information Retrieval algorithm and Heuristics.* Springer. Netherland. 2004

[4] Kang, Dae-Ki. *A Recursive Naïve Bayes Learner for Sequence Classification.* Artificial Intelligence Research Laboratory, Department of Computer Science, Lowa university. USA. 2007

[5] Khodra, Leyla Masayu dan Wibisono, Yudi. *Clustering Berita Berbahasa Indonesia.* FPMIPA Universitas Pendidikan Indonesia. 2005. http://matematika.upi.edu/staff/yudi/KNSI_*Clustering*_yudi_masayu.pdf diakses tanggal 5 Juni 2007

[6] Kim Sang-Bum dan CheolSeo, Hee et al. *Poisson Naïve Bayes for Text Classifcation with Feature Weighting.* Department of CSE.,KoreaUniversity 5-ka Anamdong, SungPuk-ku. Seoul. 2002

[7] Mandala, Rila : *Pengaruh Pembobotan Kata Pada Search Engine*. The First Conference on Telematics System. Service and Application B-6. 2004

[8] Martha, Taufiq Leo. *Penerapan Relevan Feedback Berbasis Konsep Untuk Pencarian Dokumen Pada Sistem Temu Kembali Informasi* . Institut Teknologi Sepuluh Nopember, Jurusan Teknik Informatika, Fakultas Teknologi Informasi. Surabaya. 2002

[9] Ridha, Ahmad et al. *Pengindeksan Otomatis Dengan Istilah Tunggal Untuk Dokumen Berbahasa Indonesia.* Seminar Nasional Ilmu Komputer dan Teknologi Informasi V. Departemen Ilmu Komputer, FMIPA. 2004

[10] Sentosa, Budi : *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Graha Ilmu. Surabaya. 2007

[11] Subagja, Rulan . *Aplikasi Pencarian Kata Berimbuhan Pada Kamus Bahasa Sunda Dengan Menggunakan Algoritma Stemming.* Universitas Islam Negeri, Fakultas Sains dan teknologi, Program Studi Teknin Informatika. Jakarta. 2007

[12] Susanto, Sylvia : *Klasifikasi Artikel Berita Berbahasa Indonesia Dengan Naive Bayes Classifier.* Fakultas Ilmu Komputer, Universitas Indonesia. 2006

[13] Wibisono,Yudi. *Klasifikasi Berita Menggunakan Naive Bayes Classifier.* Jurusan Pendidikan Matematika, FPMIPA, Universitas Pendidikan Indonesia. Bandung. 2005