TeknoIS: Jurnal Ilmiah Teknologi Informasi dan Sains
Volume 15 Number 1, January 2025 Page. 104-110
Journal Homepage: https://teknois.unbin.ac.id/JBS/
Received: November 2024; Accepted: January 2025; Published: January 2025

Multinomial Naïve Bayes Optimization with Information Gain for Library Book Classification

Esti Mulyani[1*], Munengsih Sari[2], Fauzan Ishlakhuddin[3]
[1,2,3]Jurusan Teknik Informatika Politeknik Negeri Indramayu
[1]Email: estimulyani@polindra.ac.id
[2]Email: munengsihsb85@polindra.ac.id
[3]Email: fauzan@polindra.ac.id
*)*Corresponding Author*

ABSTRACT

This research explores optimization in book classification activities at Politeknik Negeri Indramayu Library. The method commonly used by librarians to classify books is Dewey Decimal Classification (DDC). The DDC method allows librarians to classify book universally and systematically. However, it takes more effort and time to obtain a book classification label. This is not efficient, considering a large number of books in the library. For this reason, we propose an automatic book classification model using the text mining method. Based on the previous research, book classification model using the Multinomial Naïve Bayes (MNB) method has been conducted. The results of these studies indicate an accuracy value of 65.4%. However, the accuracy value still depends on the number of features of the dataset, so that the greater the number of features, the smaller the accuracy value. In this study, Information Gain (IG) method is proposed to select the features of a dataset in the pre- processing stage. MNB accuracy measurements are carried out based on before and after feature selection. 10-fold cross-validation is used to validate the classification model. The results showed an increase in the accuracy of MNB by 6.6%.

Keywords*: book, classification, information gain, multinomial naïve bayes.*

A. INTRODUCTION

The library is one of a learning source organization form which consists of a process of planning, organizing, moving, and controlling in a work unit to collect, store and maintain a collection of library materials that are managed and arranged systematically in a certain way by utilizing human resources to be utilized as a source of information [1]. Libraries based on their types consist of national libraries, public libraries, special libraries, school/madrasah libraries, and university libraries.

The existence of a library is a means to support the process of forming an intelligent society. In the other hands, libraries have a strategic position in the learning community because libraries are tasked with collecting, managing, and providing records of knowledge to read and study. Libraries can develop tasks properly if existing library materials can be organized and how to store them regularly, making it easier for users to get back the information needed.

Libraries have the main task in the processing of library materials, i.e. classifying books according to certain ways. Dewey Decimal Classification (DDC) is the method most widely used in the world to determine book classification (labeling) in libraries. The advantages of this DDC method are universal and more systematic [2]. DDC is a hierarchical classification system based-on the "decimal" principle to divide all fields of knowledge into 10 main classes [3]. The 10 main classes are given a code / numeric symbol (hereinafter referred to as notation). In DDC, the more specific the subject of a book, the longer the notation, because a lot of numbers are added to the basic

TeknoIS: Jurnal Ilmiah Teknologi Informasi dan Sains
Volume 15 Number 1, January 2025 Page. 104-110
Journal Homepage: https://teknois.unbin.ac.id/JBS/
Received: November 2024; Accepted: January 2025; Published: January 2025

notation. The division of book categories using DDC is carried out from general categories to specific categories.

However, book classification activities using the DDC method require a great deal of effort. Librarians must go through several stages to be able to determine the classification of books. In the first stage, the librarian can do the classification by looking at the information in the book catalog. Nevertheless, not all books have complete information in the book catalog. If the book catalog information is incomplete, then the second stage the librarian can classify by looking at the book title. However, if the book title is too general, then in the third stage the librarian can classify the book by looking at the book synopsis and analyzing it. From the analysis process, the book classification results are obtained based on the DDC method. This is less efficient considering the large number of books that must be classified in a library, as well as the labeling that must be updated following the label updates on the DDC. An automatic classification system will be the perfect solution to this problem.

Automatic classification can be done by applying the text mining method. Research on classification with text mining methods has been carried out by many previous researchers, including research that focuses on automating the final assignment classification using a short description compared to two algorithms, K-Nearest Neighbor (K-NN) and Naïve Bayes Classification, so the Naïve Bayes Classifier has the highest accuracy (65.4%) compared to K-NN (51.14%) [3]. However, this study uses the Naïve Bayes Classifier for single label classification and not multi-label classification as in the application of DDC labeling. For the case of multi-label classification, several methods have been tested to solve the related problem. Zhang [4], used multi-label K-nearest neighbor (ML-kNN) to classify web page categories.

Another study focuses on multi-label classification in holy Al-Qur'an verses using Multinomial Naïve Bayes as classifier, as well as with several stages of data preprocessing such as case folding, tokenization, and stemming. The test results in this study were to produce the best hamming loss value of 0.1247 [5].

In this study, it is proposed to apply feature selection carried out at the preprocessing stage, this is done to overcome problems in the Naïve Bayes standard classifier, namely when the features (number of words in the document, or book title) in a document are too many, the calculation results will be too small so that it cannot represented by standard floating point programming variables data types such as float or double. The application of feature selection in this study is expected to improve the performance of the Multinomial Naïve Bayes method in the library book classification process.

B. METHOD

The proposed method in this study is IG+MNB, which stands for Integration of the Information Gain (IG) feature selection applied to a book dataset which is then classified using the Multinomial Naive Bayes (MNB) algorithm. IG+MNB is proposed to achieve better classification performance than MNB.

1. MULTINOMIAL NAÏVE BAYES

The multinomial model is designed to determine the frequency of terms ie the number of times the term occurs in the document [6]. Given the fact that a term may be of great importance in determining document sentiment, the nature of this model makes it a viable choice for document classification. Apart from that, the frequency of the term also helps in deciding whether or not the term is useful in our analysis [7]. Sometimes, a term can be present in the document many times which increases the frequency of the term in this model but at the same time, it can also be a potentially meaningless keyword in the document but has a high frequency of terms, so the word -the word must be deleted first to get better accuracy of this algorithm [8][9].

Multinomial Naive Bayes is a supervised learning method, so each data needs to be labeled before training. The probability of a document d being in class c can be calculated using Equation (1).

TeknoIS: Jurnal Ilmiah Teknologi Informasi dan Sains
Volume 15 Number 1, January 2025 Page. 104-110
Journal Homepage: https://teknois.unbin.ac.id/JBS/
Received: November 2024; Accepted: January 2025; Published: January 2025

$$P\ (c|d)\ \alpha\ P(c)\ \prod_{k=1}^{n} P\ (tk|c) \quad (1)$$

P (c | d) : Probability of document d being in class c

P (c) : Prior probability of a document being in class c

{t1,.. tn}: Tokens in document d that are part of the vocabulary with number n

P(tk | c) : Conditional probability of term tk being in document of class c

Document classification aims to determine the best class for a document. The best class in Naive Bayes classification is determined by finding the maximum a posteriori (map) of a class through Equation (2).

$$Cmap = arg\ max\ P\ (c|d) = arg\ max\ P(c)\ \prod_{k=1}^{n} P\ (tk|c) \quad (2)$$

The probability value of a document in a class is obtained from the result of multiplying the prior probability value of each class with the probability value of the term in the document of a class which will then determine the highest multiplication value as the best class. To find the prior probability value of each class, we can use Equation (3).

$$P(c) = \frac{Nc}{N} \quad (3)$$

P (c) : Prior probability of each class

Nc : Number of classes in question

N : Total number of classes

Meanwhile, to find the probability value of terms in documents of a class can use Equation (4).

$$P\ (tk|c) = \frac{count\ (tk|c)+1}{count\ (c)+|V|} \quad (4)$$

2. INFORMATION GAIN

The objective of the IG is to select a subset of relevant features (words) for use in making a book classification model [10]. The use of this method can reduce the feature dimensions by measuring the Entropy reduction before and after separation. IG is also known as Mutual Information (MI) in the case of knowing the dependency between two variables (x, y). IG can be formulated as follows:

$$IG(c,t) = S(c) + \sum_{j \in value(t)}^{n} \frac{|c_j|}{|c|} S(c_j) \quad (5)$$

S(c) : entropy of all c features (before splitting),

$S(c_j)$: entropy of c feature for class t = j (after splitting),

value(t): set of possible values for t class,

n : the number of possible values for t class,

$|c_j|$: number of sample classes with value = j,

|c| : number of samples for all classes

3. DEWEY DECIMAL CLASSIFICATION

DDC method is the most widely used in the world to determine the books category (labeling) in libraries [2]. Book labels on the DDC are constantly being updated to allow for better discovery of all topics in multiple languages. The advantages of this DDC method are that it is universal and more systematic [2]. DDC is a hierarchical classification system that adheres to the

TeknoIS: Jurnal Ilmiah Teknologi Informasi dan Sains
Volume 15 Number 1, January 2025 Page. 104-110
Journal Homepage: https://teknois.unbin.ac.id/JBS/
Received: November 2024; Accepted: January 2025; Published: January 2025

"decimal" principle to divide all fields of knowledge into 10 main classes. The 10 main classes are given a code / numeric symbol (hereinafter referred to as notation). There are 10 main classes in the grouping of book categories in DDC i.e.:

Table 1. DDC Main Class

Class	*Category*
000	Computer science, information & general works
100	Philosophy & psychology
200	Religion
300	Social sciences
400	Language
500	Science
600	Technology
700	Arts & recreation
800	Literature
900	History & geography

4. PROPOSED METHOD

Figure 1 shows the proposed method framework in this research. It consists of 5 stages, i.e. 1) preprocessing the book dataset that changing the unstructured data form into structured data according to the classification needs, 2) feature extraction, 3) feature selection using IG, 4) implementing the MNB classification method for classification, and 5) the last is the measurement of the proposed method, i.e. IG + MNB.

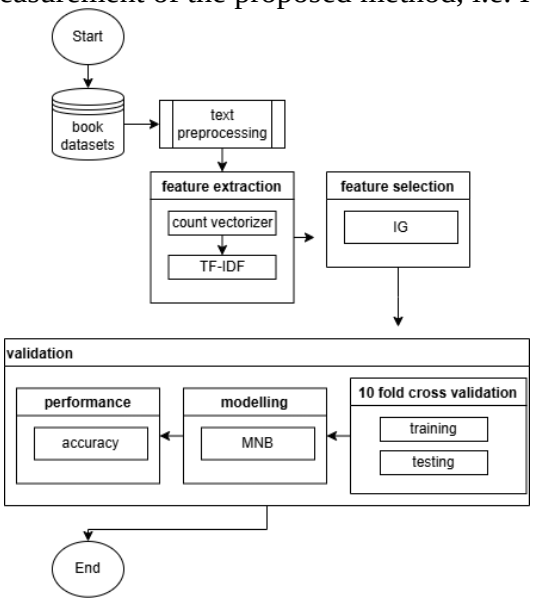

Figure 1. Proposed Method

C. RESULT AND DISCUSSION

1. PREPROCESSING

Text preprocessing in this experiment using the Natural Language Tool Kit (NLTK) Library. NLTK is a very powerful python library for use in human language data processing. Figure 2 shows the part of book dataset.

TeknoIS: Jurnal Ilmiah Teknologi Informasi dan Sains
Volume 15 Number 1, January 2025 Page. 104-110
Journal Homepage: https://teknois.unbin.ac.id/JBS/
Received: November 2024; Accepted: January 2025; Published: January 2025

	judul	klasifikasi
	title	classification
0	Red Hat Linux Networking and System Adiministr...	0
1	Voice & Data Communications Handbook Fifth Edi...	0
2	Human-Computer Interaction Third Edition	0
3	The Art of Multiprocessor Programming	0
4	Computer Networks a Systems Approach	0
5	Intel Microprocessors : Architecture	0
6	Handbook of Wireless Local Area Networks	0
7	Building The Mobile Internet	0
8	The 8051 Microcontroller	0
9	Synchronization Algorithms and Concurent Progr...	0

3387

Figure 2. Book Dataset

Experiments were conducted using a book dataset consisting of 514 English book titles consisting of 2 (two) attributes, i.e. title and classification. The book dataset can be seen in Figure 2.

In the preprocessing stage, several steps are carried out, i.e. the conversion of text into a standard form (case folding), then the process of word solving (tokenizing), and filtering (stopword removal) to retrieve important words generated from the previous process. Finally, the stemming process to remove the suffix and prefix in each word, so that a word that has a suffix or prefix will return to its basic form [11]. We can see the preprocessing result in Figure 3.

```
0     [red, hat, linux, network, system, adiministr,...
1         [voic, data, communic, handbook, fifth, edit]
2                  [humancomput, interact, third, edit]
3                       [art, multiprocessor, program]
4                   [comput, network, system, approach]
Name: judul_stem, dtype: object
```

Figure 3. Preprocessing Result

2. FEATURE EXTRACTION

At this stage, feature extraction of the preprocessing book titles is carried out [12]. Feature extraction is carried out by applying 2 methods, i.e. CountVectorizer and TF-IDF. CountVectorizer functions are used to calculate the frequency of words in book titles. CountVectorizer can turn text features into a vector representation. Meanwhile, TF-IDF or word weighting is a scheme used to calculate the weight of each word that is most commonly used.

```
Show TFIDF sample ke-0

red hat linux networking system adiministration third edition

                        TF              IDF             TF-IDF          Term

array position 53       0.125000        7.246107        0.905763        adiministration
array position 298      0.125000        2.053150        0.256644        edition
array position 418      0.125000        6.552960        0.819120        hat
array position 522      0.125000        6.147494        0.768437        linux
array position 615      0.125000        5.636669        0.704584        networking
array position 751      0.125000        6.552960        0.819120        red
array position 870      0.125000        4.538057        0.567257        system
array position 902      0.125000        4.027231        0.503404        third
```

Figure 4. Feature Extraction Result

TeknoIS: Jurnal Ilmiah Teknologi Informasi dan Sains
Volume 15 Number 1, January 2025 Page. 104-110
Journal Homepage: https://teknois.unbin.ac.id/JBS/
Received: November 2024; Accepted: January 2025; Published: January 2025

3. FEATURE SELECTION

At this stage, important and relevant features (words) are selected as well the irrelevant features (words) are reduced [10]. The Information Gain (IG) method is used in this research for selecting the features. IG uses a scoring technique to weight a feature using maximum entropy. The selected feature is a feature with an IG value that greater than or equal to a certain threshold value.

5. MODELLING AND VALIDATION

The discussion section presents the findings logically, linking them with relevant reference sources. [Times New Roman, 11, normal].

Finally, to verify whether there is a significant difference between the MNB and the proposed IG+MNB methods, the results of the two methods were compared. We tested using 10-fold cross-validation on the MNB and IG+MNB methods. The test results show the accuracy of the IG + MNB method is better than the MNB method in book classification, as in Table 2.

Table 2. Comparison of Accuracy

Accuracy	
MNB	***IG + MNB***
67,8 %	74,4%

D. CONCLUSION

Feature selection method is proposed to improve the performance of the classification algorithm. IG feature selection is applied to select important and relevant features (words) to data and reduce irrelevant features (words). The test results show that the IG + MNB method achieves a higher classification accuracy of 6.6%. Therefore, it can be concluded that the proposed method can improve the performance of the classification algorithm in the book labelling process.

E. REFERENCES

[1] Ibrahim A, *Pengantar Ilmu Perpustakaan dan Arsiparis*. Jakarta: Gunadarma Ilmu, 2017.

[2] Watthananon, "The relationship of text categorization using Dewey Decimal Classification Techniques," Int. Conf. ICT Knowl. Eng, Jan. 2015.

[3] Alalyani and Marie-Sainte, "NADA: New Arabic dataset for text classification," *Int. J. Adv. Comput. Sci. Appl*, vol. 9, no. 9, pp. 206–212, 2018.

[4] A. Frisilya and W. Yunanto, "Klasifikasi Kompetensi Tugas Akhir Secara Otomatis Berdasarkan Deskripsi Singkat Menggunakan Perbandingan Algoritma K-NN dan Naive Bayes," *Jurnal Aksara Komputer Terapan*, 2016.

[5] Pane, Mubarok, Huda, and Adiwijaya, "A multi-lable classification on topics of Quranic verses in English translation using multinomial naive bayes," Conf. Inf. Commun. Technol. ICoICT, 2019.

[6] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," Int. Conf. Autom. Comput. Technol. Manag. ICACTM, 2019.

[7] M. Abbas, K. Ali Memon, and A. Aleem Jamali, "Multinomial Naive Bayes Classification Model for Sentiment Analysis," *IJCSNS Int. J. Comput. Sci. Netw. Secur*, vol. 19, no. 3, 2019.

[8] R. Ghaniy and K. Sihotang, "Penerapan Metode Naïve Bayes Classifier Untuk Penentuan Topik Tugas Akhir," *Teknois : Jurnal Ilmiah Teknologi Informasi dan Sains*, vol. 9, no. 1, pp. 63–72, Sep. 2019, doi: 10.36350/jbs.v9i1.7.

[9] D. J. Lubis and A. I. Ningtiyas, "Penerapan Metode Naïve Bayes Untuk Rekomendasi Pemilihan Asisten Laboratorium Komputer Di Perguruan Tinggi," *Teknois : Jurnal Ilmiah Teknologi Informasi dan Sains*, vol. 12, no. 2, pp. 127–138, Jul. 2022, doi: 10.36350/jbs.v12i2.138.

[10] C. Yin and J. Xi, "Maximum entropy model for mobile text classification in cloud computing using improved information gain algorithm," *Multimed. Tools Appl*, vol. 76, no. 16, 2017.

[11] S. Kannan, "Preprocessing Techniques for Text Mining," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, 2015.

[12] M. Z. Asghar, A. Khan, S. Ahmad, and F. M. Kundi, "A Review of Feature Extraction in Sentiment Analysis Muhammad," *J. Basic. Appl. Sci. Res*, vol. 4, no. 3, 2014.