Book Classification System Based on Dewey Decimal Classification by Multinomial Naïve Bayes Method

Esti Mulyani[a,*], Munengsih Sari Bunga[b], Fauzan Islakhuddin[c], Kastuti[d]

[a,b,c,d] Politeknik Negeri Indramayu, West Java, Indonesia

Abstract

Libraries have the main task of processing library materials by classifying books according to certain methods. Dewey Decimal Classification (DDC) is the most widely used method in the world to determine book classification in libraries. However, the classification process using DDC is inefficient because it takes a long time for the large number of books in the library. This is a serious problem experienced by all libraries, so a solution is needed to bridge the problem. automatic classification system can be the right alternative to overcome the problem. In this research, an automatic classification system based on DDC using the Multinomial Naive Bayes Method so that it can speed up the classification process. This system was created using the CodeIgniter framework with the PHP programming language and MariaDB. Test results from 100 training data and 30 test data show that there are 24 test data with correct classification results and 6 test data with incorrect classification results. So it can be concluded that the accuracy rate of the test is 80%.

Keywords: Dewey Decimal Classification; Book Classification System; Multinomial Naïve Bayes; Automatic Classification; CodeIgniter Framework

1. Introduction

The library is a work unit in an educational institution. The library is a place to store a collection of library books to support the educational process. The existence of a library as a place to develop information and knowledge managed by an educational institution, as well as an educational tool to help smooth the horizons of educators and students in teaching and learning activities. Over time, libraries have also added or adapted many formats or technologies in their role of organizing and preserving information (Herron et al., 2017).

Libraries have the main task in processing library materials, namely classifying books according to certain methods. Dewey Decimal Classification (DDC) is the most widely used method in the world to determine book classification (labelling) in libraries (Watthananon, 2015). DDC is a hierarchical classification system that adheres to the "decimal" principle to divide all fields of science into 10 (ten) main classes. The 10 (ten) main classes are given a code/number symbol (hereinafter referred to as notation). In DDC, the more specialized a book subject is, the longer the notation. DDC continues to evolve over time and many structural changes occur new knowledge. This research uses the 23rd edition of the DDC structure as a reference for determining the type of book (Saha & Hatua, 2021).

Book classification activities using the DDC method require a lot of effort, librarians must go through several stages to determine book classification. In the first stage, the librarian can do the classification by looking at the information in the book catalog. However, not all books have complete information in the book catalog. If the book catalog information is incomplete, then the second stage librarians can classify by looking at the book title. However, if the book title is too general, then the third stage of the librarian can classify the book by looking at the synopsis of the book and analyzing it. From this analysis process, the results of book classification based on the DDC method are obtained. The classification process is less efficient because it takes a long time for the large number of books in the library. This is a serious problem experienced by all libraries, so a solution is needed to bridge the problem. automatic classification system can be the right alternative to overcome these problems (Mulyani et al., 2021)

Automatic classification can be done by applying text mining methods. In previous research, the classification of Turkish language documents was carried out using the Naïve Bayes approach which is one of the machine learning methods (Baygin, 2019). In other research, text mining classification can also use multinomial naive bayes (Setianingrum et al., 2018). The multinomial model is designed to determine the frequency of terms i.e. the number of times a term occurs in a document. Given the fact that a term may be very important in determining the sentiment of a document, the nature of this model

makes it a viable option for document classification. Based on the results of the description, this research developed an automatic book classification system based on DDC by applying the multinomial naive bayes method.

2. Literature Review

2.1 Library

A library is a form of learning resource organization that consists of a process of planning, organizing, supervising in a work unit and to collect, store, and maintain a collection of library materials that are systematically managed and arranged in a certain way by utilizing human resources to be used as a source of information (Ibrahim, 2017). Libraries can carry out their duties well if library materials can be organized and stored regularly, making it easier for users to retrieve the information needed.

2.2 Book Classification

Book classification is the process of grouping books according to the category they belong to. A book can be grouped into certain categories based on the words in the title of the book. Words or sentences contained in a book have certain meanings and can be used as a basis for determining the category of the book (Watthananon, 2015).

2.3 Text Category

Classification or categorization is the process of finding models or functions that explain or distinguish concepts or classes of data, with the aim of being able to estimate the class of an object whose label is unknown. Text categorization is important and the need for it will increase over time, because the data will increase over time and the storage media will also get bigger. So it is necessary to study methods for text categorization and test them through experimenting with several categorization methods.

In text categorization, a set of categories (labels) and a collection of text data are given that serves as training data, which is the data used to build the model, and then a process is carried out to find the right category for the text data, which is the data used to determine the accuracy of the model.

The core of a classification system is the pattern discovery stage, but the complete text categorization process is divided into 3 main stages, namely the initial process of text (text preprocessing), text transformation into an intermediate form (text transformation/feature generation), and pattern discovery, as can be seen in Figure 1. The initial input of this process is a text data and produces output in the form of patterns as a result of interpretation (Manning et al., 2008).
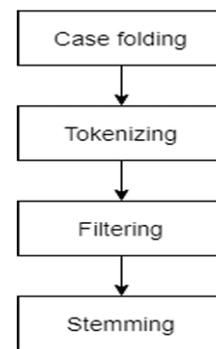

Figure 1. Text Categorization Stages

2.4 Multinomial Naïve Bayes

One of the models of Naive Bayes that is often used in text classification is Multinomial Naïve Bayes. 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).

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

P (c | d) : Probability of document d being in class c
P (c) : Prior probability of a document being in class c
{t1,t2,t3,...,tn} : Tokens in document d that are part of the vocabulary with number n
P(tk | c) : Probabilitas bersyarat term tk berada di dokumen pada kelas 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).

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

Cmap : maximum a posteriori (map) of a class

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).

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

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).

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

P (tk|c)   : The probability of term t being in a document of class c

count (tk|c) : Number of occurrences of term t in documents with class c

count (c)   : Number of occurrences of all terms in class c

|V| : Number of occurrences of all terms in all classes

2.5 Dewey Decimal Classification (DDC)

DDC is the world's most widely used method for categorizing (labeling) books in libraries. Book labels in DDC are constantly updated to enable better discovery across all topics in multiple languages. The advantages of this DDC method are universal and more systematic (Sri Eka Putri, 2017). DDC is a hierarchical classification system that adheres to the "decimal" principle to divide all fields of science into 10 (ten) main classes. The 10 (ten) main classes are given a code or numerical symbol (hereinafter referred to as notation). There are ten (10) main classes in the grouping of book categories in the DDC, consisting of:

  a) Computer Science, Information & General Works
  b) Philosophy & Psycology
  c) Religion
  d) Social Sciences
  e) Language
  f) Science
  g) Technology
  h) Arts & Recreation
  i) Literature
  j) History & Geography

3. Method

3.1 Multinomial Naïve Bayes Method

This research method generally aims to build a Web System Development for Book Type Classification Based on Dewey Decimal Classification by Utilizing the Multinomial Naive Bayes (MNB) Method using the CodeIgniter framework and PHP as a programming language. The MNB method is applied in this web system to facilitate the automatic book classification process in a library. The application of the MNB method goes through several stages: 1) dataset collection, 2) data splitting, 3) text processing, 4) MNB method. Flowchart of the implementation method can be seen from the display in Figure 2.
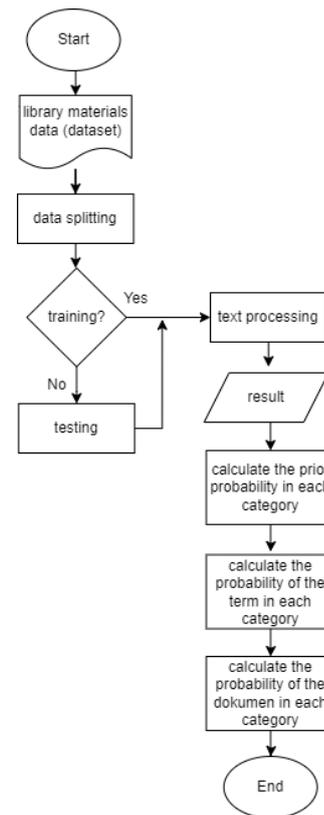

Figure 2. MNB Methods

3.1.1 Dataset Collection

The dataset needed to create this application is book data with categories based on the DDC format. For data needs, the author conducted an interview with the librarian of the Indramayu State Polytechnic. The title of the book in the dataset is in Indonesian. The list of main classes in the DDC can be seen in table 1.

Table 1. DDC Main Class List

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

3.1.2 Data Splitting

At this stage, the training data and test data are separated. The book data used as training data in the MNB method calculation process is 100 data divided into 10 data. While the test data is 30 data divided into 3 data for each main DDC class including classes 000, 100, 200, 300, 400, 500, 600, 700, 800, and 900.

3.1.3 Text Processing

In the text processing stage, text is broken down into a form that can be digested by machine learning algorithms using the MNB Method. Text processing is done through the stages of case folding, tokenizing, filtering and stemming

3.1.4 Implementation method MNB

The implementation of the MNB method is carried out with the PHP programming language using the CodeIgniter framework. The stages of this method consist of: 1) calculate the prior probability in each category, 2) calculate the probability of the term in each category, 3) calculate the probability of the document in each category.

3.2 System Architecture

The development of a Web System for Book Type Classification Based on Dewey Decimal Classification by Utilizing the Multinomial Naive Bayes Method aims to improve the services of a library in classifying the type of book collection. With this application, it is expected that rhe book classification process can be done automatically while still following the Dewey Decimal Classification (DDC) rules, making it easier for librarians to classify books and search for books as needed. The system overview can be seen from the display in Figure 3.
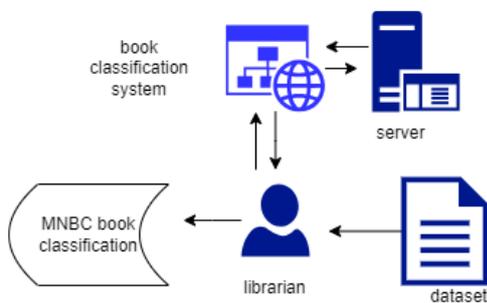

Figure 3. System Architecture

3.3 System Design

Use case diagram is a graphical representation of some or all actors, use cases and interactions between these components that introduce an application to be built, used to explain how the steps should be done by the system.
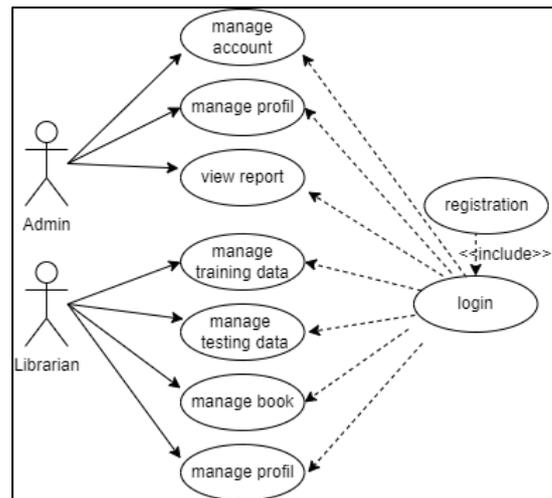

Figure 4. System Design

4. Result

This research produces a Web System Development for Book Type Classification Based on Dewey Decimal Classification by Utilizing the Multinomial Naive Bayes Method which can be used to carry out the classification process automatically with the method applied so as to improve services in a library in terms of classifying library materials in it. In this application there is only one user interface, namely Admin. Admin is in charge of processing all data including training data, test data, and book data. Admin can add, change, and delete these data.

4.1 Design Implementation
4.1.1 Login Page

The login page is used by admins and librarians to access the functions provided by the application, the display of this page can be seen in Figure 5.
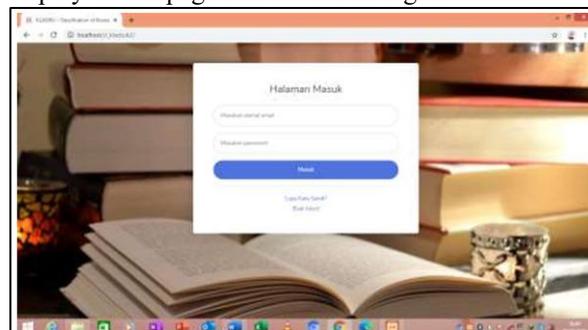

Figure 5. Login Page

4.1.2 Dashboard Page

The librarian dashboard page will appear when the librarian successfully logs in to the application, This page has some information including the amount of Training Data, Test Data, and Book Data. Display This page can be seen in Figure 6
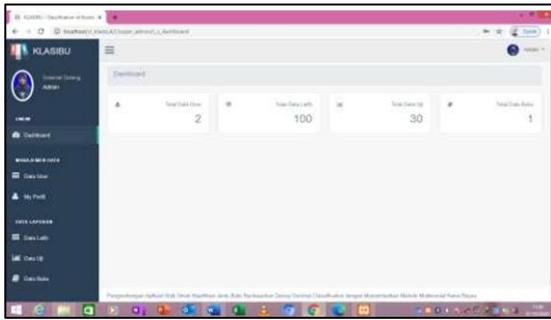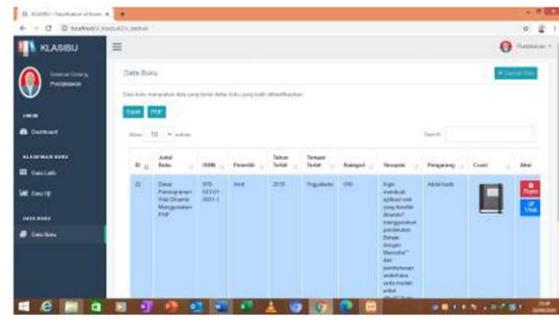
Figure 6. Dashboard Page

4.1.3 Training Page

The training data page contains a list of training data that has been added along with the results of the text processing that occurs in it. added along with the results of the text processing that occurs in it. The display of this page can be seen in Figure 7.
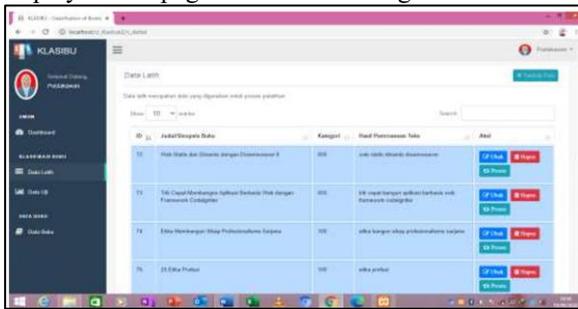

Figure 7. Training Data Page

4.1.4 Testing Data Page

The test data page contains a list of test data that has been added along with the results of the text processing and classification process that occurs in it. The display of this page can be seen in Figure 8.
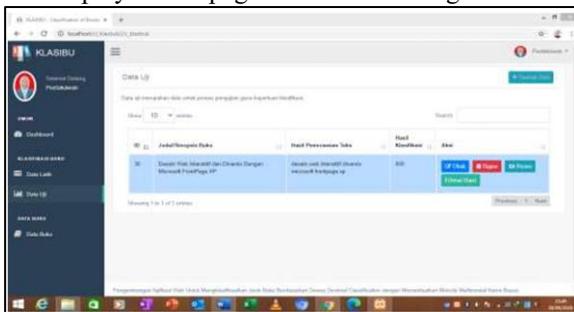

Figure 8. Testing Data Page

4.1.5 Book Data Page

The book data page contains a list of book data that has been added. This book data list can not only be can also be changed and deleted. The display of this page can be seen in Figure 9.

Figure 9. Book Data Page

4.2 Method Implementation

In the application of the MNB method, testing was carried out on 30 test data with 100 training data consisting of all main classes in DDC and available along with the results of text processing. The dataset used is book data at the Indramayu State Polytechnic Library. The title of the book in the dataset is in Indonesian.

4.2.1 Training Data

Book data used as training data in MNB method calculation process consists of 100 data divided into 10 data for each main DDC class including classes 000, 100, 200, 300, 400, 500, 600, 700, 800, and 900. An example of some of the data can be seen in Table 2.

Table 2. Sample Training Data on Method Implementation

No	Book Title	Class
1	Web Statis dan Dinamis dengan Dreamweaver 8	000
2	Filsafat Timur : Sebuah pengantar ke Pemikiran-Pemikiran Timur	100
3	Mudah & Praktis Belajar Tajwid	200
4	Pendidikan Kewargaan (Civic Education) : Demokrasi Hak asasi Manusia dan Masyarakat Madani	300
5	Kamus Mandarin Indonesia 8000 Kata	400
6	Fisika Dasar	500
7	Jam Piket Organ Tubuh	600
8	101 Tip & Trik Kamera Digital	700
9	Buku Penuntun Membuat Tesis Skripsi Disertasi Makalah	800
10	Antara Daerah Dan Negara Indonesia Tahun 1950-an	900

4.2.2 Testing Data

The book data used as test data in the calculation process of the multinomial naive bayes method consists of 30 data divided into 3 data for each main DDC class including classes 000, 100, 200, 300, 400, 500, 600, 700, 800, and 900. An example of some of the data can be seen in Table 3.

Table 3. Sample Testing Data on Method Implementation

No	Book Title	Class
1	Desain Web Interaktif dan Dinamis dengan Microsoft FrontPage XP	000
2	Pola Dasar Pengembangan SDM P3JJ : Profesi dengan 3 Jalur Ber Jenjang	100
3	Dahsyatnya Membaca & Menghafal Alqur'an	200
4	Pendidikan Kewarganegaraan dan Pancasila untuk Perguruan Tinggi	300
5	Ilmuan dan Bahasa Indonesia	400
6	Mekanika Teknik Edisi keempat	500
7	Soal-Soal Biologi Lengkap Tes Masuk Perguruan Tinggi	600
8	Shalat an Amazing	700
9	Dewey	800
10	Biografi Abu Bakar Ash Shiddiq	900

4.2.3 Classification Result

The classification results of 30 test data show 24 test data with correct classification results and 6 test data with incorrect classification results (indicated by the red line). The 6 incorrect test data include books entitled "Pola Dasar Pengembangan SDM P3JJ : Profesi dengan 3 Jalur Ber Jenjang", "Soal-Soal Lengkap Biologi Tes Masuk Perguruan Tinggi", "Shalat an Amazing", "Residential Lighting : A Practical Panduan Desain Cantik dan Berkelanjutan Edisi Kedua", "Dewey", dan "Biografi Abu Bakar Ash Shiddiq". From the results of the classification, the level of accuracy or precision of the classification process carried out can be calculated using the following equation:

$$Akurasi = \frac{Hasil\ Klasifikasi\ Tepat}{Keseluruhan\ Data\ Uji}\ X\ 100\%$$

$$Akurasi = \frac{24}{30}\ X\ 100\% = 80\%$$

The calculation shows that from 30 book data used as test data in the book type classification process based on the Dewey Decimal Classification format with the multinomial naive Bayes method, it produces an accuracy level of 80%.

5. Conclusion

Development of a web system for book type classification based on dewey decimal classification by utilizing the naive multinomial method bayes method was built using the codeigniter framework with the PHP programming language and the MYSQL database system.

Automatic book classification in this application development is made by applying the DDC format to determine book categories and the multinomial naive bayes method for the classification process.

The application of the multinomial naive bayes method in the development of this application is used to perform an automatic book classification process by applying the principle of the highest probability value

of word occurrence in the title or synopsis of a book in a book category based on the DDC format.

Test results from 100 training data and 30 test data show that there are 24 test data with correct classification results and 6 test data with incorrect classification results. So it can be concluded that the accuracy rate of the test is 80%

References

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: Capitalizing on big data. Journal of Travel Research, 58(2), 175–191. https://doi.org/10.1177/0047287517747753

Aliyah, K., Kisworo, B., Miraj, S., & Gupta, S. (2024). Pengaruh media sosial terhadap pemberdayaan masyarakat. Jurnal Pendidikan dan Pemberdayaan Masyarakat, 11(1), 26–38.

Arman, A. A., & Sidik, A. P. (2019, October 19–20). Measurement of engagement rate in Instagram (Case study: Instagram Indonesian government ministry and institutions). In 2019 International Conference on ICT for Smart Society (ICISS) (pp. 1–6). Bandung, Indonesia. IEEE. https://doi.org/10.1109/ICISS48059.2019.896982 6Bloor, D., & Moffatt, S. (2017). Consumer behavior in digital marketing. Journal of Marketing Research, 54(3), 455–468.

Baygin, M. (2019). Classification of Text Documents based on Naive Bayes using N-Gram Features. *2018 International Conference on Artificial Intelligence and Data Processing, IDAP 2018*, 1–5. https://doi.org/10.1109/IDAP.2018.8620853

Freberg, K., Graham, K., McGaughey, K., & Freberg, L. A. (2011). Who are the social media influencers? A study of public perceptions of personality. Public Relations Review, 37(1), 90–92. https://doi.org/10.1016/j.pubrev.2010.11.001

Flinta, N. (2022). The evolution of interdisciplinary research methodologies. Journal of Interdisciplinary and Multidisciplinary Research Literature, 7(1), 1–15.

George, O. A., & Ramos, C. M. Q. (2024). Wellness tourism and digital transformation. International Journal of Spa and Wellness, 7(2), 139–161. https://doi.org/10.1080/24721735.2024.000000

Herron, N. L., Editor, G., Fourie, D. K., Dowell, D. R., Edward Evans, G., Intner, S. S., & Weihs, J. (2017). *Developing Library and Information Center Collections*. https://archive.org/details/developinglibraryOOedwa

Ibrahim, A. (2017). *Pengantar Ilmu Perpustakaan dan Arsiparis*. Gunadarma Ilmu.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. Business Horizons, 53(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Kumar, V., Pozza, I. D., & Ganesh, J. (2016). Multichannel marketing: Mindset and program development. Journal of Marketing, 80(2), 40–58. https://doi.org/10.1509/jm.15.0189

Liu, B. (2012). Sentiment analysis and opinion mining. Morgan & Claypool Publishers. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Manning, C. D., Raghavan, P., & Schütze, H. (2008). An introduction to information retrieval. Cambridge University Press.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Monacho, B. C., & Slamet, Y. (2023). Strategi komunikasi pemasaran digital. Jurnal Komunikasi, 39(3), 373–388.

Mulyani, E., Muhamad, F. P. B., & Cahyanto, K. A. (2021). Pengaruh N-Gram terhadap Klasifikasi Buku menggunakan Ekstraksi dan Seleksi Fitur pada Multinomial Naïve Bayes. *Jurnal Media Informatika Budidarma*, *5*(1), 264. https://doi.org/10.30865/mib.v5i1.2672

Niwareeba, R., Cox, M. A., & Cheng, L. (2023). Adaptive-mode PAPR reduction algorithm for optical OFDM systems leveraging lexicographical permutations. Electronics, 12(13), 2797. https://doi.org/10.3390/electronics12132797

P Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) (pp. 1320–1326). European Language Resources Association.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

Saha, S., & Hatua, S. R. (2021). A Comparative Analysis of Structural Layout Changes in 1st To 23rd Edition of Dewey Decimal Classification Scheme. *Library Herald*, *59*(3), 56–64. https://doi.org/10.5958/0976-2469.2021.00022.1

Setianingrum, A. H., Kalokasari, D. H., & Shofi, I. M. (2018). Implementasi Algoritma Multinomial Naive Bayes Classifier. *Jurnal Teknik Informatika*, *10*(2), 109–118. https://doi.org/10.15408/jti.v10i2.6822

Smith, A., & Zook, Z. (2011). Social media marketing strategies. Journal of Business Research, 64(3), 271–278. https://doi.org/10.1016/j.jbusres.2009.11.014

Sri Eka Putri. (2017). *Keterampilan Pustakawan dalam Mengklasifikasi* Pustaka *dengan Menggunakan Dewey Decimal Classification*.

Stolte, D. (2023). Assessing the engagement of social media marketing: The case study of a tourism Toronto video-based marketing campaign (Master's thesis, Toronto Metropolitan University). Retrieved from https://rshare.library.torontomu.ca/articles/thesis/Assessing_the_engagement_of_social_media_marketing_the_case_study_of_a_tourism_Toronto_video-based_marketing_campaign/14647614.

Toly, A. A., Angel, M., & Roedyanto, J. F. (2023). Digital conference proceedings and their impact. Journal of International Conference Proceedings, 6(1), 35–52.

Varadainy, Z., Helita, W., Hermansyah, M., & Rahmawati, I. D. (2024). Business trends in digital learning platforms. International Journal of Business, Learning, and Professional Studies, 1(1), 1–6.

De Vries, L., Gensler, S., & Leeflang, P. S. H. (2012). Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. Journal of Interactive Marketing, 26(2), 83–91. https://doi.org/10.1016/j.intmar.2012.01.003

Wang, B., Jin, X., & Ma, N. (2024). Artificial intelligence in marketing. Kybernetes. https://doi.org/10.1108/K-04-2023-0456

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). Advances in artificial intelligence: A comprehensive review. Artificial Intelligence Review, 55(8), 5731–5780. https://doi.org/10.1007/s10462-022-10164-9

Watthananon, J. (2015). The relationship of text categorization using Dewey Decimal Classification techniques. *International Conference on ICT and Knowledge Engineering*, *2015-Janua*(January), 72–77. https://doi.org/10.1109/ICTKE.2014.7001538

Yew, R. L. H., Suhaidi, S. B., Seewoochurn, P., & Sevamalai, V. K. (2018, October 26–28). Social network influencers' engagement rate algorithm using Instagram data. In 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA) (pp. 1–6). IEEE. https://doi.org/10.1109/ICACCAF.2018.8776839

Yu, X., Hua, J., Bao, T., & Zhang, H. (2023). A novel PAPR reduction scheme for low-cost terminals in 6G. Energies, 16(8), 3307. https://doi.org/10.3390/en16083307