Jurnal Pendidikan dan Konseling
Volume 5 Nomor 1 Tahun 2023
E-ISSN: 2685-936X dan P-ISSN: 2685-9351
Universitas Pahlawan Tuanku Tambusai

Use of the K-Nearest Neighbor (KNN) Algorithm in New Categories Based on Books

Annisa Helmina[1*], Fadia Kalma Lailani[2], Fatihaturahmi[3], Nizwardi Jalinus[4], Waskito[5]
[1,2,3,4,5] Faculty of Engineering - Universitas Negeri Padang, Indonesia
Email: annisahell@student.unp.ac.id[1*]

Abstrak

Di era yang terus berkembang ini, menuntut anak bangsa menjadi insan yang cerdas. Bangsa yang cerdas adalah bangsa yang harus terus belajar. Belajar membutuhkan bahan atau panduan yang akan menemani perjalanan. Hal ini membutuhkan buku sebagai sarana proses pembelajaran. Buku adalah hasil pemikiran yang berisi pengetahuan yang telah dianalisis. Untuk mendapatkan buku yang cocok untuk Anda, Anda harus mencari buku tersebut di toko buku. Namun untuk menemukan buku yang diinginkan, calon pembeli harus membaca setiap deskripsi buku dan menghabiskan banyak waktu dalam proses pencarian. Dari permasalahan tersebut, penulis mendapatkan ide untuk membuat sistem kategorisasi baru berdasarkan koleksi buku. Tujuan dari penelitian ini adalah membangun sistem klasifikasi buku untuk menghindari masalah yang melibatkan kategorisasi buku yang lebih akurat. Penelitian ini menggunakan teknik text mining dan algoritma K-Nearest Neighbor (KNN) sebagai logika pemrogramannya. Data pelatihan yang digunakan dalam sistem ini adalah 35 buku yang merupakan jumlah keseluruhan buku. Sedangkan data testing merupakan data yang memiliki persentase pencarian lebih tinggi. Berdasarkan data training yang tersedia, dengan data k=3 testing “beginner”, “youtuber” dan “study” “ melalui tahapan TF-IDF dan Cosine Similarity, hasil rekomendasi kategorisasi baru diurutkan berdasarkan nilai kemiripan data dari tertinggi hingga terendah yaitu buku. B9 adalah 12,7% , B11 adalah 11,6% dan B13 adalah 9,0%.
Kata kunci: *KNN, Text Mining, Kategori Baru, Buku*

Abstract

In this era that continues to develop, it requires the nation's children to become intelligent people. An intelligent nation is a nation that must continue to learn. Learning requires materials or guides that will accompany the journey. This requires books as a means of the learning process. Books are the result of thoughts that contain knowledge that has been analyzed. To get the book that suits you, you have to looking for the book at the bookstore. However, finding the book you want requires prospective buyers to read every description of the book and it spends a lot of time in the search process. From these problems, the author came up with an idea to create a new categorization system based on a collection of books. The purpose of this research is to build a book classification system in order to avoid problems involving more accurate book categorization. This study uses text mining techniques and the K-Nearest Neighbor (KNN) algorithm as the programming logic. The training data used in this system is 35 books which is the total number of books. While testing data is data that has a higher percentage of searches . Based on the available training data, with k=3 data testing “beginner”, “youtuber” and “study” “ through the TF-IDF and Cosine Similarity stages, the results of the new categorization recommendations are sorted by the data similarity values from the highest to the lowest, namely books. B9 is 12.7 % , B11 is 11.6% and B13 is 9.0%.
Keywords*: KNN, Text Mining, New Category, Book*

INTRODUCTION

The development of today's economic world, especially bookstores that are so mushrooming, makes competition for this business unavoidable and increasingly fierce [1] . A book is a collection of paper or other material that is bound together at one end and contains writing [2] . Information is data that has been processed into a form that is useful for the recipient and is real, in the form of value that can be understood in current and future decisions [3]. The increasing development of technology that is getting faster makes more and more kinds of books circulating on the internet [4]. Basically, books are only enjoyed physically, but now books can be enjoyed by books through electronic media devices [5]. Currently the library revolution has entered the Tulungagung area, this bookstore revolution is marked by many online bookstores such as Gramedia [6]. Printing PT. Gramedia is a company engaged in the printing industry and has been operating since 1972 [7].

Because there are many descriptions of books in bookstores, it makes it difficult for potential buyers because they have to read every available book description to get the book they want. therefore required a system that can categorize books according to the description of the book that is input. Category description of the book in this study using text mining techniques whose goal is to find the words contained in the description of the book so that it can represent the contents of the description . Text mining in the narrow sense is only a method that can form specific or easily known new information from a cluster of documents [8]. The main procedure in this method is related to finding words that can represent the contents of the document for further analysis of the relationships between documents using certain statistical methods such as group analysis, classification and association [9]. In this study the authors used the K-Nearest Neighbor Algorithm in categorizing Gramedia books .

To generate a new category in several collections of books, the author focuses on the description of the category of computer network books that are the data is taken from the online bookstore Gramedia using the KNN algorithm . The working principle of K-Nearest Neighbor (KNN) is to find the shortest distance between the data to be evaluated and its K nearest neighbors in the training data [10]. From the discussion and the existing problems, the authors are interested in taking the title , namely "Using the k-Nearest Neighbor (KNN) Algorithm in a New Category Based on a Collection of Books" with a case study of this research on Gramedia.

METHOD

This study uses qualitative methods using text mining techniques. The dataset that will be used as training data consists of 35 total books. The testing data is taken from the description data inputted by the user in the book search column.

Text Mining

Text mining is a knowledge exploration process based on certain patterns of textual data retrieval [11]. Before the process of text mining techniques, the text data is carried out first through the preprocessing stage. Preprocessing is the steps in processing the next dataset which will be included in the classification system. Here are the steps that the author did in the preprocessing stage [12]. Pre-processing of data is the process of cleaning and preparing the text for classification because at the word level, many words in the comments do not fit the general orientation [13].

The preprocessing stages are as follows:

1. Case Folding

   This stage converts a text into the same form, namely changing all letters to lowercase so that the consistency of the text is to get more optimal results [12].

2. tokenization

   Tokenization is the process of cutting sentences in a set of text documents into chunks of words or

characters that suit the system's needs and removing certain characters such as punctuation marks. These pieces are called tokens [14].

3. Filtering

   Filtering is done with the stopwords method by eliminating all conjunctions, pronouns, and others such as he, we, you, like, for, and others [15].

4. Stemming

   Stemming is a process that is used to return a word to its root word [16]or is a process of converting an affixed token to a root word, by removing all affixes in the token [17].

Term Frequency-Inverse Document Frequency (TF-IDF)

(TF-IDF) is a method used to determine how far the terms are related to the document by giving each word a weight [18].

1. TF (Term frequency)

   TF is the number of terms in a script [19]. TF is the number of terms that appear in the document. The resulting weight will be even greater if the TF occurrence value is high if the TF occurrence value is high. TF uses a comparison of the frequency of a word with the total number of words in the document. The formula is:

   $$tf = 0.5 + 0.5 \frac{(tf)}{(tfmax)} \quad (1)$$

2. IDF (Inverse Document Frequency)

   IDF is formulated with the following formula:

   $$idf = log \frac{D}{df} \quad (2)$$

   Information:

   D is the number of existing documents from d_f

   J is the number of documents in term t_j

3. The formula for TF-IDF is the multiplication between TF and IDF, as follows:

   $$Wij = tf\ x\ log \frac{D}{df} \quad (3)$$

   Where:

   w_{ij} is the weight of the term t existing documents d

   tf is the number of occurrences of term t in the document d

   D is the number of all documents in the database

   Df is the number of documents containing the term t

K-Nearest Neighbor Algorithm

K-Nearest Neighbor is one method for making decisions using supervised learning where the results of the new input data are classified based on the closest in the value data [20]. The KNN algorithm is a method that uses a supervised algorithm that belongs to the instance-based learning group and also a lazy learning technique by searching for groups of k objects in the training data that are closest (similar) to the objects in the new data or testing data [21]. KNN aims to classify objects based on attributes and training samples that are only based on memory by providing query points, it will find a number of k objects or (training points) closest to the query point [22].

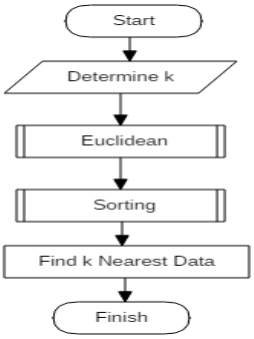

Figure 1 . K-Nearest Neighbor Algorithm Flowchart

1. Set score k
2. Calculates the distance between the evaluated data and all training data
3. Sort the results from the distance that has been formed
4. Find the closest k data
5. Collect the same or the appropriate class
6. Finds and sets the number of the closest neighboring class as the data class to be evaluated.

There are many ways to measure the distance between the closeness of the new data and the old data (training data), but what is used in this study is cosine similarity which is a method to calculate the similarity between two objects expressed in two vectors by using the keywords of a document as the size .

$$cosSom(dj, qk) = \frac{\sum_{i=1}^{n}(tdij\ x\ tqik)}{\sqrt{\sum_{i=1}^{n} tdij2\ x\ \sum_{i=1}^{n} tqik2}} \qquad (3)$$

Information :

cosSim(dj,qk) is document similarity level with certain queries .

tdij is the ith term in the vector for the jth document

tqik is the ith term in the vector for the kth query

n is the number of unique terms in the data set

To determine the similarity, it can be used with several functions, namely the similarity function, and the closeness between distances . Cosine similarity . The following is the formula for finding similarities using Euclidean Distance:

$$Similarity = 100\% - \frac{d_{(x1,x2)}}{d_{(x1,x2)}max} \times 100\% \qquad (6)$$

RESULTS AND SYSTEM DESIGN

System design in this study uses use case diagrams and activity diagrams which are described as follows:

Use Case Diagrams

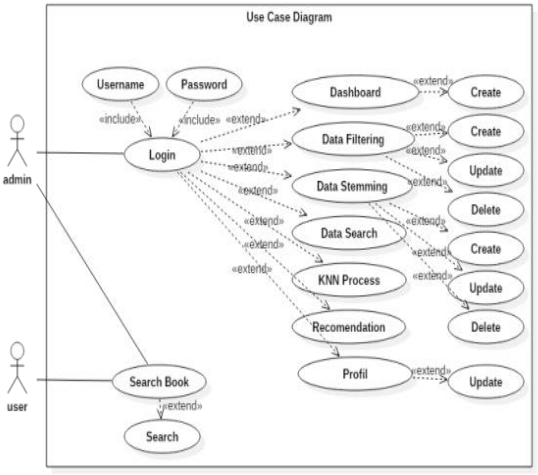

Figure 2 . Use Case Diagram Application New Categorization

The use case diagram in this system has two users, namely "user" and "admin". “user” can only search for books and see the new categorization of books. While the "admin" can manage all systems by logging in first.

System Workflow

The system workflow using the K-NN algorithm can be seen in Figure 3 below:

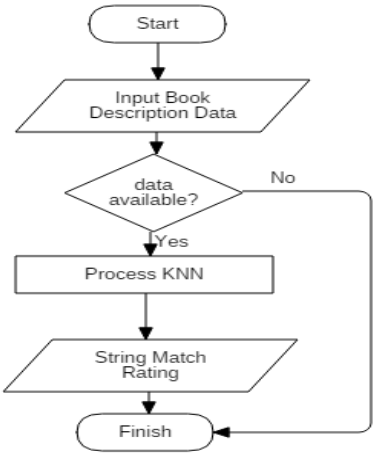

Figure 3. System workflow

The workflow of this system is to input the book description data first, then the system will see whether the data is available or not in the book data in the system. Furthermore, if it is available, then the input data will be processed through the K-NN algorithm which will then match it with the existing string. Meanwhile, if the data entered is not available, then the system will immediately provide complete information.

Before performing calculations using the K-Nearest Neighbor algorithm , the training data and testing data are first determined . From the data collection stage from the e - Gramedia website, it can be seen the process of the system running as follows:

1. Training Data

   data (training data) is a raw material for information obtained based on existing data at the research site. The book data that will be used in this manual calculation takes a sample of 35 book data in the category of computer network books on the Gramedia website .

2. Data Testing

Testing data selected to look for manual calculations is the data obtained from the most user search results . In this system, it happens that the data that users often look for manually are the words "beginner", "youtuber" and "learning".

Data testing will be processed through the pre-processing stage , so the data is ready to be processed using the K-Nearest Neighbor algorithm which is used to group data based on the closest distance/ the level of similarity of the data with the existing dataset / training data. To measure the level of similarity using the K-Nearest Neighbor algorithm , this study uses the Cosine Similarity formula to calculate the similarity/closeness of the documents . In this study, calculations using the K-Nearest Neighbor algorithm include :

1. Calculating word weight using Term Frequency–Inverse Document Frequency (TF-IDF)

Table 1 . Calculating word weight using TF

Data	Term		
	Beginner	Youtuber	Study
	TF		
T1	1	1	1
B1	0	0	1
B2	3	0	3
B3	0	0	0
B4	1	0	0
B5	0	0	0
Amount	**6**	**2**	**9**

Table 2 . Calculating word weights using IDF

Data	Term		
	Beginner	Youtuber	Study
	IDF		
T1	0.76591679	1.243038	0.5898
B1	0	0	0.5898
B2	2.29775038	0	1.7695
B3	0	0	0
B4	0.76591679	0	0
B5	0.76591679	1.243038	0.5898
Amount	**0.76592**	**1.243038**	**0.58983**

As in the table above, there are 35 book data that will search for the words "beginner", " youtuber " and "learning". So the TF-IDF calculation is as follows:

a. From the 35 book data, the word beginner appears 1 time in book B2, and is repeated in B4, B13, B24, B31, and B32. So the calculation results are as follows:

$tf = 1$ $\quad$ $idf = log\left({}^{35}/_{6}\right)$= 0.76592

$df = 6$

b. From the 35 book data, the word youtuber appears once in book B9, and repeats itself in B11. So the calculation results are as follows:

$tf = 1$ $idf = log\left({}^{35}/_{2}\right)$= 1.243038

$df = 2$

c. From the 35 data books, the word learning appears once in book B1, and is repeated in B2, B9, B13, B14, B16, B31, B33, and B35. So the calculation results are as follows:

$tf = 1$ $idf = log\left({}^{35}/_{9}\right)$= 0.58983

$df = 9$

2. Calculating the level of similarity (cosine similarity)

After the weights are obtained through searching with TF-IDF, the next step is to calculate the distance or level of similarity of the data with each existing training data using the cosine similarity distance formula . Then, the system will sort the values from the highest to the lowest distance, as follows:

Table 3. Calculation of cosine similarity

Data	Term			∑WT1*WBi
	Beginner	Youtubers	Study	
	WTI*WBi			
T1	1	1	1	0.3478942
B1	0	0	1	2.8035681
B2	3	0	3	0
B3	0	0	0	0.5866285
B4	1	0	0	0
B5	0	0	0	0.3478942

3. Calculating Vector Length

Calculate the length of the vector or the result of the square of each term in each document (including T_1) then add up and take the root . Calculate the length of the vector as follows:

Table 4. Calculation of vector length

Data	Term			Wn ²	Wn ²
	Beginner	Youtuber	Study		
	Wn ²				
T1	0.586629	1.5451436	0.347894	2.479666	1.574696
B1	0	0	0.347894	0.347894	0.589826
B2	5.279657	0	3.131047	8.410704	2.900121
B3	0	0	0	0	0
B4	0.586629	0	0	0.586629	0.765917
B5	0	0	0	0	0

4. Cosine Similarity Calculation

Doing the division between the results of WT1*WBi with $\sqrt{\sum W_{n^2}}$, then the value of cosine similarit is obtained with the formula:

$$\left(\frac{\sum WT1 * WBi}{\sqrt{\sum W_{T^2}} * \sqrt{\sum W_{B^2}}}\right) \quad (7)$$

Table 5. Calculation of Cosine Similarity

Book	Cos(T,Bi)	Percent
B1	0.374565	5.51%
B2	0.613901	9.02%
B3	0	0.00%
B4	0.48639	7.15%
B5	0	0.00%

After obtaining the Cosine Similarity value , then sorting the level of similarity of the data. From these results, the data similarity values can be sorted from the highest to the lowest, namely:

Table 6. Results Sequence of data similarity values

Book	Cos(T,Bi)	Percent
B9	0.865594	12.7%
B11	0.789383	11.6%
B13	0.613901	9.0%
B2	0.613901	9.0%
B31	0.588497	8.7%

After knowing the distance from the highest to the lowest, the highest k data will be taken. In this case, k = 3. If the chosen value of k for KNN is 3, then the 3 highest similarity values will be selected, then obtained:

Table 4.8 Calculation results k=3

k=3	
Book	**Cosine Similarity**
B9	12.7%
B11	11.6%
B13	9.0%

From the results of these calculations, it is obtained recommendations for new categories of new categorization cases in a collection of Gramedia books using the K-Nearest Neighbor algorithm. These are beginners, youtubers and learning. These words are obtained based on the most searched words by users with the highest cosine similarity for k = 3 is 12.7%.

CONCLUSION

To find out the type of book you want, of course, you have to read the description of the book one by one. However, this will take a long time so we need a system that can categorize books from several collections of books available. The system that was created uses the K-NN algorithm as its programming logic. This study has 35 books taken from the Gramedia website to be used as training data with data

limitations only on computer network books. In this system, 3 words that are often entered by the user in the search are used as testing data. This system also uses text mining techniques in the search process. Based on the available training data, with k=3 data testing “beginner”, “youtuber” and “learning” “ through the TF-IDF and Cosine Similarity stages, the results of the new categorization recommendations are sorted by the data similarity values from the highest to the lowest, namely books. B9 is 12.7 % , B11 is 11.6% and B13 is 9.0%.

REFERENCES

E. Y. Bulele, “Analisis Pengaruh Citra Toko, Kualitas Pelayanan Dan Ketersediaan Produk Terhadap Keputusan Pembelian Di Toko Buku Gramedia Manado,” *J. Berk. Ilm. Efisiensi*, vol. 16, no. 3, pp. 258–269, 2016.

B. Santoso, “Perancangan Aplikasi Olap (Online Analitycal Processing) Penjualan Buku Pada Toko Buku Gramedia Lubuklinggau Dengan Metode Clustering,” *J. Teknol. Inf. MURA*, vol. 9, no. 2, pp. 40–42, 2017.

E. A. Lisangan and N. T. S. Saptadi, “Perancangan Data Warehouse Pengolahan Persediaan Buku PT. Gramedia Asri Media Makassar,” *Semin. dan Call Munas Aptikom*, pp. 81–90, 2010.

M. Maulidah, Windu Gata, Rizki Aulianita, and Cucu Ika Agustyaningrum, “Algoritma Klasifikasi Decision Tree Untuk Rekomendasi Buku Berdasarkan Kategori Buku,” *E-Bisnis J. Ilm. Ekon. dan Bisnis*, vol. 13, no. 2, pp. 89–96, 2020.

R. P. Sutanto, “Studi Kasus Website Gramedia sebagai Media Online untuk Membeli Buku,” *Nirmana*, vol. 17, no. 1, p. 37, 2018, doi: 10.9744/nirmana.17.1.37-41.

F. Rozi, F. Sukmana, and M. N. Adani, “Pengelompokkan Judul Buku dengan Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Term Frequency – Inverse Document Frequency (TF-IDF),” *JIMP J. Inform. Merdeka Pasuruan*, vol. 6, no. 3, pp. 1–5, 2021.

G. Gunadi and D. I. Sensuse, “Penerapan Metode Data Mining Market Basket Analysis Terhadap Data Penjualan Produk Buku Dengan Menggunakan Algoritma Apriori Dan Frequent Pattern Growth (Fp-Growth) : Studi Kasus Percetakan PT. Gramedia,” *Telematika*, vol. 4, no. 1, pp. 118–132, 2012.

A. Ririd, P. Y. Saputra, and ..., “Sistem koreksi kesalahan pengetikan kata kunci dalam pencarian artikel menggunakan algoritma jaro-winkler,” *Semin. Inform.*, pp. 60–65, 2019.

M. P. R. Putra and K. R. N. Wardani, “Penerapan Text Mining Dalam Menganalisis Kepribadian Pengguna Media Sosial,” *JUTIM (Jurnal Tek. Inform. Musirawas)*, vol. 5, no. 1, pp. 63–71, 2020.

Jumaidi, “Sistem Pendukung Keputusan Untuk Menentukan,” *J. Istek*, vol. VI, no. 1, pp. 40–42, 2013.

G. A. Nursanto, I. A. Prabadhi, and A. A. A. Pramana, “Implementasi Text Mining Untuk Pengklasifikasian Pertanyaan Masyarakat Melalui Whatsapp Dengan Metode Naïve Bayes Classifier (Studi Kasus: Kantor Imigrasi Kelas I Khusus TPI Surabaya),” *Temat. | Technol. Manag. Informatics Res. Journals*, vol. 5, no. 1, pp. 43–66, 2021.

M. F. Trisnadi, S. Al Faraby, and M. Dwifebri, “Sentiment Analysis pada Movie Review Menggunakan Feature Selection Mutual Information dan K-Nearest Neighbour Classifier,” *e-Proceeding Eng.*, vol. 8, no. 5, pp. 1–11, 2021.

E. Indrayuni, “Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes,” *J. Khatulistiwa Inform.*, vol. 7, no. 1, pp. 29–36, 2019.

M. Christianto, J. Andjarwirawan, and A. Tjondrowiguno, “Aplikasi analisa sentimen pada komentar berbahasa Indonesia dalam objek video di website YouTube menggunakan metode Naïve Bayes classifier,” *J. Infra*, vol. 8.1, pp. 255–259, 2020.

E. Septrinas, Indriati, and A. A. Soebroto, “Klasifikasi Berita Olahraga Berbahasa Indonesia menggunakan Metode BM25 dan K-Nearest Neighbor,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 10, pp. 9762–9769, 2019.

E. R. Setyaningsih, “Sentiment Classification untuk Opini Berita SepakBola,” *J. Intell. Syst. Comput.*, vol. 3, no. 2, pp. 93–98, 2021.

R. T. Adek, M. Fikry, and A. Helmina, “Opinion Mining About Parfum on E-Commerce Bukalapak.Com Using the Naïve Bayes Algorithm,” *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 6, no. 1, pp. 107–

114, 2020.
B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, “Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity,” *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 1, pp. 306–312, 2018.
L. Tommy, C. Kirana, and V. Lindawati, “Recommender System Dengan Kombinasi Apriori Dan Content-Based Filtering Pada Aplikasi Pemesanan Produk,” *J. Teknoinfo*, vol. 13, no. 2, p. 84, 2019.
J. Pseudocode, V. I. I. I. Nomor, S. Kasus, D. Pemuda, and P. Bengkulu, “Implementasi Metode K-Nearest Neighbor (Knn) Dan Simple Additive Weighting (Saw) Dalam Pengambilan Keputusan Seleksi Penerimaan Anggota Paskibraka,” *J. Pseudocode*, vol. III, no. 2, pp. 98–112, 2016.
L. Data *et al.*, “Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa,” *J. Sains, Teknol. dan Ind.*, vol. 13, no. 2, pp. 195–202, 2016.
S. Ernawati and R. Wati, “Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel,” *J. Khatulistiwa Inform.*, vol. VI, no. 1, pp. 64–69, 2018.