ISSN 2723-6471


# Recommender System for Book Review based on Clustering Algorithms

Devi Udariansyah[1*], Tri Basuki Kurniawan[2], Deshinta Arrova Dewi[3],
Mohd Zaki Zakaria[4], Nur Syuhana binti Abd Hanan[5]

[1]*Faculty Science Technology, Universitas Bina Darma, Palembang, Indonesia*

[2]*Postgraduate Program, Universitas Bina Darma, Palembang, Indonesia*

[3]*Faculty of Data Science and Information Technology, INTI International University, Malaysia*

[4,5]*Faculty of Computer and Mathematics Sciences, University Technology Mara, Malaysia*

(Received: June 22, 2024; Revised: September 20, 2024; Accepted: October 17, 2024; Available online: December 28, 2024)

**Abstract**

Book reviews show the expression of the reviewers that are to be evaluated and describe the book. Today, the amount of the book is growing rapidly, and it offers people a lot of choices. The recommender system on book reviews is mostly mentioned, and we will recommend a book based on the keyword selected. This study highlights two primary objectives. The first objective is to identify the keywords of the book review, and the last objective is to design and develop a book review analysis visualization using the result of the k-means clustering algorithm. The methodology of this research consists of ten phases, which start with the preliminary study, knowledge acquisition and analysis phase, data collection phase, data pre-processing phase, and modeling phase. The research then continues with the design and implementation, dashboard development, testing and evaluation, and finally, the documentation phase. The data from this study is scraped from Amazon.com and focuses on three genres: Fiction and Fantasy, Mystery and Thriller, and Romance. All the data will be clean before it can be applied to k-means clustering. The result of clustering will define the keywords for every genre and will compare with the keywords for each book that was collected from Amazon.com.

*Keywords:* Book Review, Recommender System, Clustering Algorithms, Process Innovation

## 1. Introduction

A book is a medium for documenting information in the form of writing or images, typically made up of several pages bound together and covered by a cover. For example, the books contain knowledge or histories [1]. Why should we read books? Reading is good for our brain, as it increases blood flow and improves brain connectivity [2]. The most important thing when reading a book is it will increase the knowledge of history or culture and also increase the skill in an area of interest. It will make it easier to handle any problem and give more confidence. People can buy a book easily at the bookstore or website shop. If purchased on the website, people can place their rating and review on the website that they were purchased. Nowadays, we have not only a book in physical but also an electronic book, as known as eBooks [3]. An electronic book, or eBook, is a digital version of a printed book that includes various types of material, such as text, figures, and pictures, which can be read on computers, eReaders, tablets, and smartphones [4].

Reviews of items are important as they lead others to use or do not impact production profit and popularity. Users of different perspectives have written a summary text. Useful visual analysis of online customer reviews, which has a big impact on the analysis of reviews, has generated a high level of interest in determining effective decision-making [5].In this work approach, book reviews' analysis visualization is created and implemented, which helps to recommend books based on the keyword that gets from the result of clustering.

Recommendations were obtained from reviews and ratings from users to find a list of books [6]. For the visualization, it will use a clustering approach. Clustering is based on similarity, where similar elements are kept in a single group. The idea of dividing large data into smaller sets seemed to offer advantages, such as scalability, which, due to the smaller set of data on which algorithms operate, improves response time. Clustering has been widely investigated as

***Corresponding author: Devi Udariansyah (devi.udariansyah@binadarma.ac.id)**
**DOI: https://doi.org/10.47738/jads.v6i1.492**

an unsupervised method [7]. People have a lot of options in the real world, and the main challenge is selecting one out of them [6]. The person who does not have enough personal experience to evaluate often takes the other person's opinion in choosing one of many items. But, by taking their opinion, it will take a long time to make a decision. Moreover, the increase in the amount of digital content available and the number of Internet users has created an information problem that discourages timely access to items of interest on the Internet [8].

These days, many books have been published, and many genres of books can be found. Whenever the user searches for a book online, he/she gets confused by the number of items shown [6]. Therefore, this research aims to solve these problems by developing a visualization to show the keyword that users mostly give reviews in certain books.

## 2. Literature Review

A recommender system is a straightforward algorithm designed to provide users with the most relevant information by identifying patterns within a dataset. This algorithm rates items and highlights those with high ratings, suggesting products that best align with consumer preferences [9]. One of the key features of a recommender system is its ability to predict user interests and needs by analyzing their behavior or the behavior of other users, ultimately creating a personalized recommendation experience [10].

Recommender systems can determine whether a particular shopper might be interested in a specific product by analyzing the customer's profile [11]. These systems function by collecting user inputs, applying suitable algorithms, and generating recommendations tailored to the user's choices and preferences [12]. Implementing recommender systems is a smart solution for streamlining the search process on E-commerce websites, saving users time, and simultaneously increasing profits by encouraging consumers to explore and purchase more products [5].

The rapid development of mobile technology has resulted in a large influx of information available to users. While this is a positive trend, it raises an urgent challenge: how to deliver sufficient and relevant information efficiently. Fortunately, extensive research has been conducted to address this issue [13]. For example, Lian incorporated the matrix factorization model and spatial clustering phenomena in humans to improve recommendation efficiency [14]. Similarly, Zhang introduced a hybrid recommendation system that enhanced user recommendations to a notable degree [5]. Interestingly, Kacchi developed a client-server model where users act as clients submitting queries to the system [15]. Recommender systems have become indispensable tools in the E-commerce environment, effectively managing information overload. Over the past decade, these systems have been widely adopted by major platforms such as Amazon.com, Netflix, and others, contributing to increased sales by utilizing diverse recommendation techniques [16], [17].

## 3. Methodology

The research process begins with a structured methodology, which includes the following phases: preliminary study, knowledge acquisition, data collection, data pre-processing, system architecture design, system development, system testing and evaluation, and documentation. Each phase is designed to address and achieve the objectives of the study. The research methodology serves as a comprehensive summary of the phases and activities undertaken in this study. Careful consideration is given to selecting the appropriate methods to ensure that the study's results align with the stated objectives outlined in the main proposal. The methodology representation is clearly explained, with outcomes evaluated to verify their alignment with the research objectives. Figure 1 provides a visual representation of the research methodology framework, illustrating the logical flow of the study's phases and their interconnections.

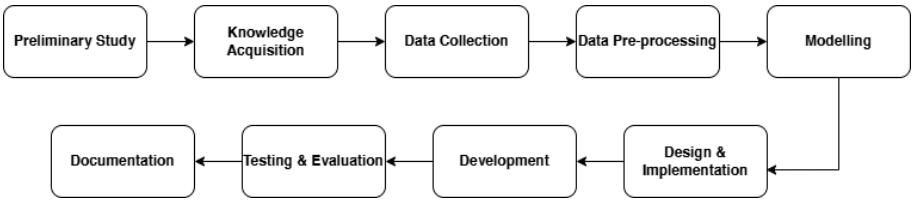



**Figure 1.** Research Methodology Framework

### 3.1. Preliminary Study

The preliminary study involves periodic discussions about the research and an in-depth exploration of the problem statement. This phase includes examining the domain and the problem statement through various resources such as the

 ISSN 2723-6471 

Internet, academic journals, forums, books, and other academic writings authored by researchers relevant to the topic. These reviews are instrumental in refining the study's title, problem statement, research questions, objectives, and scope. The primary challenge in the domain of book recommendation lies in the overwhelming number of books available, which makes it difficult for readers to decide which book to choose. The study also investigates existing methods used in this domain. Based on this exploration, the researcher opted to adopt a straightforward approach for recommending books, focusing on book reviews as the key determinant in the recommendation process.

## 3.2. Knowledge Acquisition

This phase is to gain the information and knowledge collected from the literature review about the domain along with the problems and techniques that are related and possible solutions to the domain problem. For this study, journal articles or books are reviewed to find the best technique for designing the website and to set the constraints. This phase represents an important step towards achieving a good understanding of the study concept collected in the literature review and research methodology. All of the relevant techniques are compared and selected to suit the needs of the study to produce the best results. Text clustering and k-mean clustering are the best techniques to solve this problem. Every technique has its advantages and disadvantages. The techniques will refer to the data that has been obtained and explained in the subtopic in this chapter.

## 3.3. Data Collection

The data collection phase outlines the process of identifying and gathering relevant data for the study. During this phase, we explored methods for extracting data from websites. Data from Amazon.com was collected using Python-based tools, specifically Scrapy and Beautiful Soup, which were customized to meet the requirements of the study [18]. Web scraping refers to the process of extracting or scraping data from websites and is sometimes called web data collection or web data extraction. In its simplest form, web scraping involves copying and pasting text from a website to a local device. However, with tools such as web crawlers, web scraping can automate the collection of data. Web crawlers are scripts that connect to the World Wide Web via the Hypertext Transfer Protocol (HTTP) to retrieve and extract data automatically [19]. To extract structured data from websites, Beautiful Soup is employed. Beautiful Soup is a Python library specifically designed for parsing HTML and XML files. It simplifies interactions with HTML structures and serves as a support module for extracting required data efficiently [20], [21].

The focus of this study is on three genres: Science Fiction and Fantasy, Mystery and Thriller, and Romance. The first step in the data extraction process is to list book titles along with their details. Beautiful Soup is used for this purpose due to its user-friendly capabilities in web scraping. The key data fields extracted include Book Name, Author, Rating, Number of Customer Ratings, and Price. Figure 2 illustrates the Amazon websites used for data collection. Figure 3 and figure 4 display the code utilized for web scraping and its continuation, respectively.

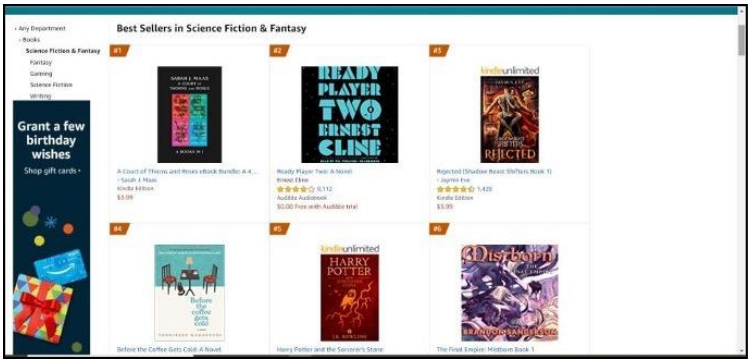



**Figure 2.** Amazon websites that are used

 ISSN 2723-6471

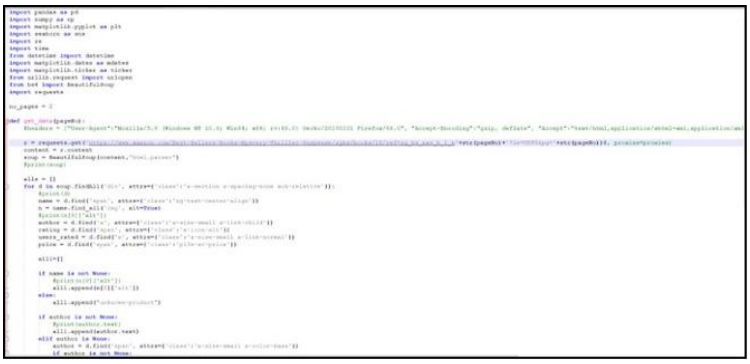



**Figure 3.** Coding for scraping

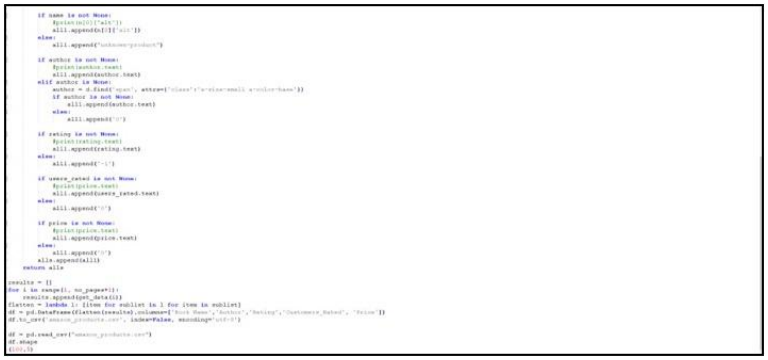
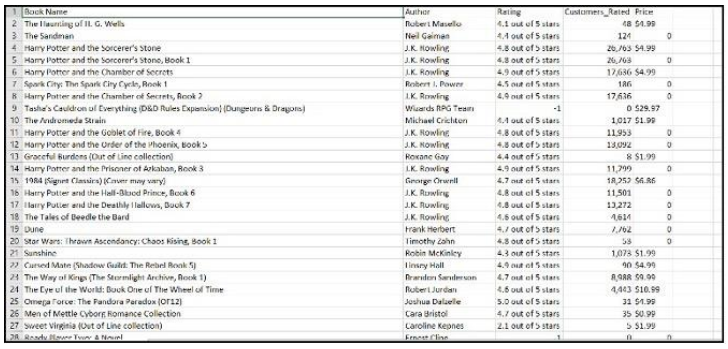


**Figure 4.** Continuation of coding for scraping

Figure 3 and figure 4 show the code that will perform the following functions. First, the function get_data inside a for_loop will be called, starting from 1 to the number of pages+1 the for loop will iterate over this function. Since the output is a nested list, the list will be compressed first and then passed to the DataFrame. Finally, the CSV file saves the data frame, as shown in figure 5 and table 1.



| | Book Name | Author | Rating | Customers_Rated | Price | |
|---|---|---|---|---|---|---|
| 2 | The Haunting of H. G. Wells | Robert Masello | 4.1 out of 5 stars | 48 | $4.99 | |
| 3 | The Sandman | Neil Gaiman | 4.4 out of 5 stars | 124 | | 0 |
| 4 | Harry Potter and the Sorcerer's Stone | J.K. Rowling | 4.8 out of 5 stars | 26,763 | $4.99 | |
| 5 | Harry Potter and the Sorcerer's Stone, Book 1 | J.K. Rowling | 4.8 out of 5 stars | 26,763 | | 0 |
| 6 | Harry Potter and the Chamber of Secrets | J.K. Rowling | 4.9 out of 5 stars | 17,636 | $4.99 | |
| 7 | Spark City: The Spark City Cycle, Book 1 | Robert J. Power | 4.5 out of 5 stars | 186 | | 0 |
| 8 | Harry Potter and the Chamber of Secrets, Book 2 | J.K. Rowling | 4.9 out of 5 stars | 17,636 | | 0 |
| 9 | Tasha's Cauldron of Everything (D&D Rules Expansion) (Dungeons & Dragons) | Wizards RPG Team | -1 | 0 | $29.97 | |
| 10 | The Andromeda Strain | Michael Crichton | 4.1 out of 5 stars | 1,017 | $1.99 | |
| 11 | Harry Potter and the Goblet of Fire, Book 4 | J.K. Rowling | 4.8 out of 5 stars | 11,953 | | 0 |
| 12 | Harry Potter and the Order of the Phoenix, Book 5 | J.K. Rowling | 4.8 out of 5 stars | 13,092 | | 0 |
| 13 | Graceful Burdens (Out of Line collection) | Roxane Gay | 4.4 out of 5 stars | 8 | $1.99 | |
| 14 | Harry Potter and the Prisoner of Azkaban, Book 3 | J.K. Rowling | 4.9 out of 5 stars | 11,799 | | 0 |
| 15 | 1984 (Signet Classics) (Cover may vary) | George Orwell | 4.7 out of 5 stars | 18,252 | $6.86 | |
| 16 | Harry Potter and the Half-Blood Prince, Book 6 | J.K. Rowling | 4.8 out of 5 stars | 11,501 | | 0 |
| 17 | Harry Potter and the Deathly Hallows, Book 7 | J.K. Rowling | 4.8 out of 5 stars | 13,272 | | 0 |
| 18 | The Tales of Beedle the Bard | J.K. Rowling | 4.6 out of 5 stars | 4,614 | | 0 |
| 19 | Dune | Frank Herbert | 4.7 out of 5 stars | 7,762 | | 0 |
| 20 | Star Wars: Thrawn Ascendancy: Chaos Rising, Book 1 | Timothy Zahn | 4.8 out of 5 stars | 58 | | 0 |
| 21 | Sunshine | Robin McKinley | 4.3 out of 5 stars | 1,073 | $1.99 | |
| 22 | Cursed Mate (Shadow Guild: The Rebel Book 5) | Linsey Hall | 4.9 out of 5 stars | 90 | $4.99 | |
| 23 | The Way of Kings (The Stormlight Archive, Book 1) | Brandon Sanderson | 4.7 out of 5 stars | 8,988 | $9.99 | |
| 24 | The Eye of the World: Book One of The Wheel of Time | Robert Jordan | 4.6 out of 5 stars | 4,443 | $10.99 | |
| 25 | Omega Force: The Pandora Paradox (OF12) | Joshua Dalzelle | 5.0 out of 5 stars | 31 | $4.99 | |
| 26 | Men of Mettle Cyborg Romance Collection | Cara Bristol | 4.7 out of 5 stars | 35 | $0.99 | |
| 27 | Sweet Virginia (Out of Line collection) | Caroline Kepnes | 2.1 out of 5 stars | 5 | $1.99 | |
| 28 | Ready Player Four: A Novel | Ernest Cline | | 1 | | 0 |

**Figure 5.** Data after scraping

**Table 1.** Number of books that are scraped

| Genre | Number of books | Genre |
|---|---|---|
| Science fiction and fantasy | 100 books | Science fiction and fantasy |
| Mystery and thriller | 100 books | Mystery and thriller |
| Romance | 100 books | Romance |

Table 1 above shows the number of books scraped from Amazon based on genre. The total scraped reviews are 300 books. The total of books will change due to missing value on reviews after removing the books that do not have reviews and reviews less than 200 by using Excel. The reason more than 200 reviews are chosen is because we want at least 50 books for every genre. So, it will have many reviews that can be used for clustering. Next, table 2 shows the number of selected books.

ISSN 2723-6471


**Table 2.** Number of selected books

| Genre | Number of books | Genre |
|---|---|---|
| Science fiction and fantasy | 50 books | Science fiction and fantasy |
| Mystery and thriller | 62 books | Mystery and thriller |
| Romance | 50 books | Romance |

Table 2 shows the updated number of books, and the list of books will be used to scrape the reviews for each book. The next step is to scrape the reviews of the book by using Scrapy. For large-scale web scraping, Scrapy is a Python framework. It was written in Python, and it provides all the resources that we need to extract data from websites effectively and process and store it in a suitable structure and format. The reviews will be scraped based on the title shown in figure 6 and reviews in figure 7.


**Figure 6.** Title of book

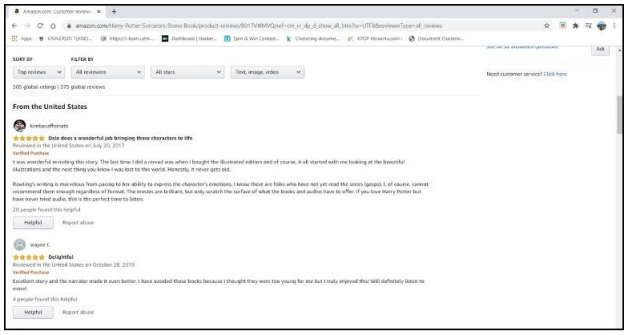



**Figure 7.** Reviews of book

The important data is comment since it will be used in clustering. It would need to do some cleaning after the data has been scraped before it can be used to cluster. The data collected from scraping is scattered and not accurate. The keywords that are mentioned in reviews from the Amazon website are also scraped since we want to compare them with the keywords from the result of clustering. Figure 8 shows the keywords for each book, and figure 9 shows the list of keywords that have been scraped and saved as .csv.

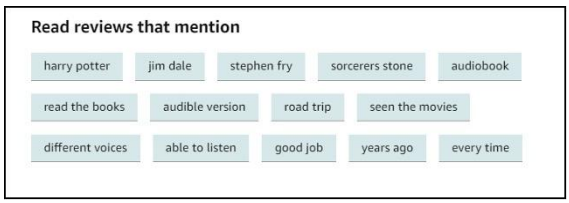



**Figure 8.** Keywords that are mentioned in reviews

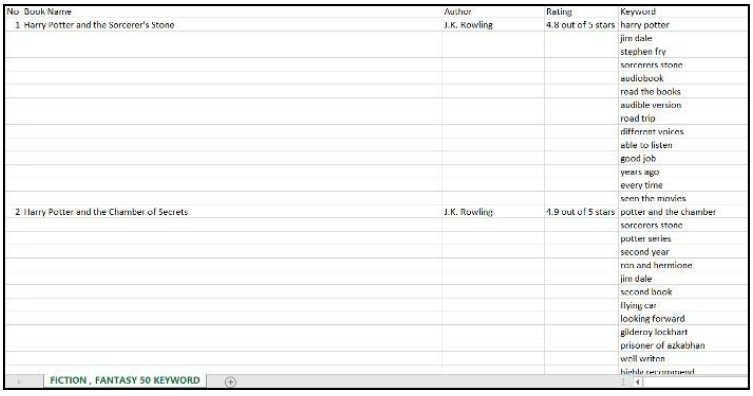



**Figure 9.** Keyword for genre Fiction and Fantasy

Pre-processing of data is a technique to turn raw data into a convenient and efficient format. Data pre-processing phases include data cleaning, data transformation, and data reduction. Data cleaning is about handling data with irrelevant and missing data. Missing data can be handled in a variety of ways, for example, ignoring the tuples or filling in the missing values, while it can be handled in the following ways for noisy data such as binning method, regression, and clustering. This process also removes a symbol number and makes it into lowercase. Figure 10 shows the code that is used for data cleaning.

 ISSN 2723-6471


```
#Data cleaning
import re
first_text=df.Review[0]
text=re.sub("[^a-zA-Z]"," ",first_text) #changing characters with space
text=text.lower()
```

**Figure 10.** Code for removing symbols and making it lowercase

Stopword removal is applied during the data cleaning phase to eliminate words that do not contribute significant meaning to a sentence. Examples of such stopwords include "this," "is," "can," "do," "more," and "such." These are identified and removed using the Natural Language Toolkit (NLTK) stopword collection. This step ensures that the tokenized vector retains only meaningful words, enhancing the quality of the processed data.

Next, data transformation is performed to prepare the data for the mining process. This involves steps such as normalization, attribute selection, and discretization. Additionally, data reduction techniques are applied to manage large datasets effectively and remove duplicate entries, ensuring a streamlined dataset for further analysis.

In the modeling phase, clustering techniques are employed to compare the clustering results and evaluate their performance. For this study, k-means clustering was chosen because it is well-suited for unlabeled data, where no predefined categories or groups exist. K-means clustering is a type of unsupervised learning algorithm that aims to partition the data into K distinct groups based on the given features. The algorithm assigns each data point to one of the K groups based on feature similarity. To visualize the clustering, Principal Component Analysis (PCA) is used, as it effectively captures the global structure of the data. PCA is a dimensionality-reduction technique that condenses a large set of variables into a smaller, more manageable set while retaining significant variance.

In addition to k-means, Silhouette Analysis is used to evaluate the quality of clustering and identify keywords from each cluster. A silhouette score close to +1 indicates that a data point is well-separated from neighboring clusters. A score of 0 suggests the data point lies on the boundary between clusters, while a negative score indicates potential misassignment to the wrong cluster.

When designing the system, a well-defined system architecture, including the system cycle and flowchart, is crucial. These components are developed using open-source tools based on the knowledge acquired during the study. The system architecture provides a comprehensive representation of the entire study, covering processes from web scraping to the deployment of the main platform. It culminates in generating the final book recommendations, as illustrated in figure 11.

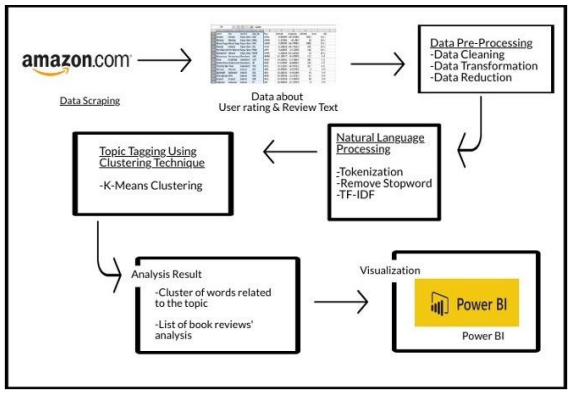



**Figure 11.** System Architecture

The system flow begins with the scraping phase, where Scrapy was utilized to collect book reviews. The scraped data then undergoes several pre-processing steps, including data cleaning, data transformation, and data reduction. Using k-means clustering, keywords are extracted from the book reviews and subsequently visualized on a dashboard developed with Power BI. These processes are explained in detail, and the step-by-step flow is illustrated in figure 12. During the system development phase, a dashboard was designed and developed to visualize the clustering results. This provides a clear and intuitive way to interpret the keywords from the clusters of book reviews. Additionally, the dashboard allows users to search for book titles based on the extracted keywords from the reviews, facilitating a more efficient exploration process. Figure 12 presents an example of the dashboard to be developed. The resources utilized in this phase include journals, books, and websites to gain a thorough understanding of the methods applied.

 ISSN 2723-6471 

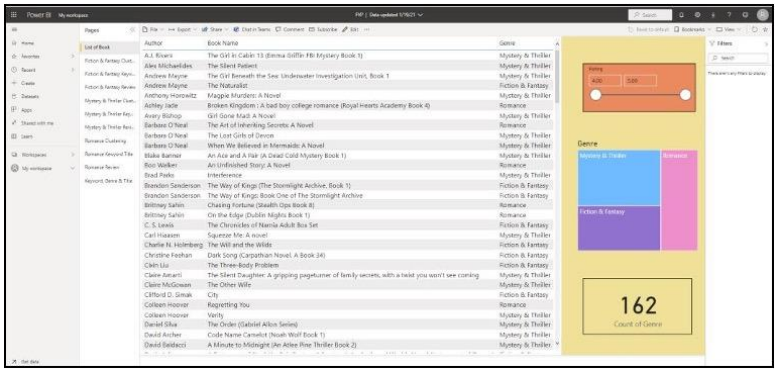



**Figure 12.** Dashboard using Power BI

The next phase is the system testing and evaluation. System testing is required for each study to make sure the filtering and sorting are used correctly. Based on some selected criteria, filters in Power BI sort data and information. Different fields or values can be selected within the fields, and only the relevant details can be displayed. We have determined that the dashboard can display the information and make users easily understand it. The dashboard needs to be examined in various circumstances to see the effectiveness of the proposed dashboard.

## 4. Results and Discussion

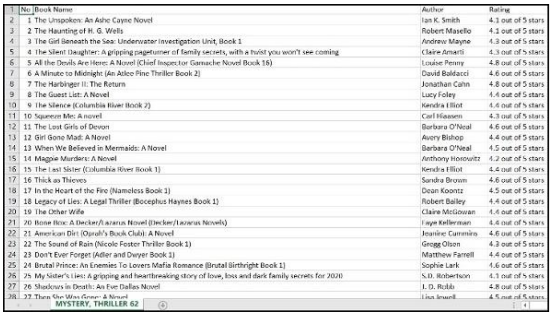
This chapter documents the results of the study and presents the findings obtained during its development. The results and discussion focus on key aspects, including data collection, data analysis and interpretation, data visualization, and the overall outcomes derived from the study. Each section is designed to provide a detailed examination of the processes and insights gained during the research. The data collection results highlight the extracted data after pre-processing, where books without reviews and reviews with fewer than 200 words were removed to ensure data quality and relevance. This refinement step ensures that the analysis is based on meaningful and comprehensive reviews. Following this, the cleaned dataset is categorized based on genres, enabling the study to focus on specific preferences within the Fiction and Fantasy, Mystery and Thriller, and Romance genres.

Figure 13 presents the titles of books from the Fiction and Fantasy category that remained after applying the review filtering process. A total of 28 book titles were selected as samples, offering a representative overview of this genre. This dataset serves as the foundation for further analysis, such as identifying patterns, trends, and key themes present in the reviews. Figure 14 focuses on the Mystery and Thriller genre, providing a similar analysis. The filtered books in this category offer insights into user preferences, common themes, and popular titles within this genre. These results help illustrate how clustering and keyword extraction can refine large datasets to identify actionable insights.

| No | Book Name | Author | Rating |
|---|---|---|---|
| 1 | Harry Potter and the Sorcerer's Stone | J.K. Rowling | 4.8 out of 5 stars |
| 2 | Harry Potter and the Chamber of Secrets | J.K. Rowling | 4.8 out of 5 stars |
| 3 | The Andromeda Strain | Michael Crichton | 4.4 out of 5 stars |
| 4 | Harry Potter and the Goblet of Fire, Book 4 | J.K. Rowling | 4.8 out of 5 stars |
| 5 | Harry Potter and the Order of the Phoenix, Book 5 | J.K. Rowling | 4.8 out of 5 stars |
| 6 | Harry Potter and the Prisoner of Azkaban, Book 3 | J.K. Rowling | 4.9 out of 5 stars |
| 7 | 1984 (Signet Classics) (Cover may vary) | George Orwell | 4.7 out of 5 stars |
| 8 | Harry Potter and the Half-Blood Prince, Book 6 | J.K. Rowling | 4.8 out of 5 stars |
| 9 | Harry Potter and the Deathly Hallows, Book 7 | J.K. Rowling | 4.8 out of 5 stars |
| 10 | The Tales of Beedle the Bard | J.K. Rowling | 4.6 out of 5 stars |
| 11 | Dune | Frank Herbert | 4.7 out of 5 stars |
| 12 | Sunshine | Robin McKinley | 4.3 out of 5 stars |
| 13 | The Way of Kings (The Stormlight Archive, Book 1) | Brandon Sanderson | 4.7 out of 5 stars |
| 14 | The Eye of the World: Book One of The Wheel of Time | Robert Jordan | 4.6 out of 5 stars |
| 15 | Transfusion: A Vampire King Paranormal Romance (Transfusion Saga Book 1) | Stephanie Hudson | 4.3 out of 5 stars |
| 16 | The Stand | Stephen King | 4.7 out of 5 stars |
| 17 | Player's Handbook (Dungeons & Dragons) | Wizards RPG Team | 4.8 out of 5 stars |
| 18 | The Chronicles of Narnia Adult Box Set | C. S. Lewis | 4.8 out of 5 stars |
| 19 | Fahrenheit 451 | Ray Bradbury | 4.6 out of 5 stars |
| 20 | Emergency Skin (Forward collection) | N. K. Jemisin | 4.2 out of 5 stars |
| 21 | The Promise of Hades: A Fated Mates Fantasy Romance (The Hades Trials Book 3) | Eliza Raine | 4.8 out of 5 stars |
| 22 | Lovecraft Country: A Novel | Matt Ruff | 4.6 out of 5 stars |
| 23 | The Fellowship of the Ring: Book One in The Lord of the Rings Trilogy | J. R. R. Tolkien | 4.8 out of 5 stars |
| 24 | World War Z: The Complete Edition: An Oral History of the Zombie War | Max Brooks | 4.8 out of 5 stars |
| 25 | The Enigma Cube (Alien Artifact Book 1) | Douglas E. Richards | 4.3 out of 5 stars |
| 26 | The Glamourist (The Vine Witch Book 2) | Luanne G. Smith | 4.6 out of 5 stars |
| 27 | Dark Song (Carpathian Novel: A Book 14) | Christine Feehan | 4.7 out of 5 stars |

FICTION , FANTASY 50

**Figure 13.** List of books for genre Fiction and Fantasy



| No | Book Name | Author | Rating |
|---|---|---|---|
| 1 | The Unspoken: An Ashe Cayne Novel | Ian K. Smith | 4.1 out of 5 stars |
| 2 | The Haunting of H. G. Wells | Robert Masello | 4.1 out of 5 stars |
| 3 | The Girl Beneath the Sea: Underwater Investigation Unit, Book 1 | Andrew Mayne | 4.3 out of 5 stars |
| 4 | The Silent Daughter: A gripping page-turner of family secrets, with a twist you won't see coming | Claire Ansari | 4.3 out of 5 stars |
| 5 | All the Devils Are Here: A Novel (Chief Inspector Gamache Novel Book 16) | Louise Penny | 4.8 out of 5 stars |
| 6 | A Minute to Midnight (An Adler Pine Thriller Book 2) | David Baldacci | 4.6 out of 5 stars |
| 7 | The Harbinger II: The Return | Jonathan Cahn | 4.8 out of 5 stars |
| 8 | The Guest List: A Novel | Lucy Foley | 4.4 out of 5 stars |
| 9 | The Silence (Columbia River Book 2) | Kendra Elliot | 4.4 out of 5 stars |
| 10 | Squeeze Me: A novel | Carl Hiaasen | 4.3 out of 5 stars |
| 11 | The Lost Girls of Devon | Barbara O'Neal | 4.6 out of 5 stars |
| 12 | Girl Gone Mad: A Novel | Avery Bishop | 4.4 out of 5 stars |
| 13 | When We Believed in Mermaids: A Novel | Barbara O'Neal | 4.5 out of 5 stars |
| 14 | Magpie Murders: A Novel | Anthony Horowitz | 4.2 out of 5 stars |
| 15 | The Last Sister (Columbia River Book 1) | Kendra Elliot | 4.4 out of 5 stars |
| 16 | Thick as Thieves | Sandra Brown | 4.6 out of 5 stars |
| 17 | In the Heart of the Fire (Nameless Book 1) | Dean Koontz | 4.5 out of 5 stars |
| 18 | Legacy of Lies: A Legal Thriller (Bocephus Haynes Book 1) | Robert Bailey | 4.4 out of 5 stars |
| 19 | The Other Wife | Claire McGowan | 4.4 out of 5 stars |
| 20 | Bone Box: A Decker/Lazarus Novel (Decker/Lazarus Novels) | Faye Kellerman | 4.4 out of 5 stars |
| 21 | American Girl (Oprah's Book Club): A Novel | Jeanine Cummins | 4.6 out of 5 stars |
| 22 | The Sound of Rain (Nicole Foster Thriller Book 1) | Gregg Olson | 4.3 out of 5 stars |
| 23 | Don't Ever Forget (Adler and Dwyer Book 1) | Matthew Farrell | 4.4 out of 5 stars |
| 24 | Brutal Prince: An Enemies To Lovers Mafia Romance (Brutal Birthright Book 1) | Sophie Lark | 4.6 out of 5 stars |
| 25 | My Sister's Keeper: A gripping and heartbreaking story of love, loss and dark family secrets for 2020 | S.D. Robertson | 4.3 out of 5 stars |
| 26 | Shadows in Death: An Eve Dallas Novel | J. D. Robb | 4.8 out of 5 stars |
| 27 | The Girl She Was Once: A Novel | Lisa Jewell | 4.5 out of 5 stars |

MYSTERY, THRILLER 62

**Figure 14.** List of books for genre Mystery and Thriller

Figure 15 showcases the Romance genre, which is another popular category in book recommendations. The filtered dataset includes books with significant user engagement, as indicated by the presence of comprehensive reviews. This allows for an in-depth examination of reader sentiment and preferences, which are critical for building effective recommendation models.

| No | Book Name | Author | Rating |
|---|---|---|---|
| 1 | I Promise You: Stand-Alone College Sports Romance | Ilsa Madden-Mills | 4.8 out of 5 stars |
| 2 | The Haunting of Brynn Wilder: A Novel | Wendy Webb | 4.4 out of 5 stars |
| 3 | The Return | Nicholas Sparks | 4.6 out of 5 stars |
| 4 | Where the Forest Meets the Stars | Glendy Vanderah | 4.6 out of 5 stars |
| 5 | Empire High Elite | Ivy Smoak | 4.9 out of 5 stars |
| 6 | Where the Crawdads Sing | Delia Owens | 4.8 out of 5 stars |
| 7 | Playing with Fire: A Bad Boy College Romance | L.J. Shen | 4.7 out of 5 stars |
| 8 | Mr. Fixer Upper | Lucy Score | 4.7 out of 5 stars |
| 9 | Roommaid: A Novel | Sariah Wilson | 4.3 out of 5 stars |
| 10 | Single Dad Seeks Juliet: A Feel-Great Romantic Comedy | Max Monroe | 4.7 out of 5 stars |
| 11 | The Best Friend Zone: A Small Town Romance | Nicole Snow | 4.6 out of 5 stars |
| 12 | Chasing Fortune (Stealth Ops Book 8) | Brittney Sahin | 4.9 out of 5 stars |
| 13 | On the Edge (Dublin Nights Book 1) | Brittney Sahin | 4.4 out of 5 stars |
| 14 | Temptation (The Hunted Series Book 1) | Ivy Smoak | 4.3 out of 5 stars |
| 15 | Empire High Untouchables | Ivy Smoak | 4.7 out of 5 stars |
| 16 | The Dare | Lauren Landish | 4.4 out of 5 stars |
| 17 | The Art of Inheriting Secrets: A Novel | Barbara O'Neal | 4.5 out of 5 stars |
| 18 | See Her Die (Bree Taggert Book 2) | Melinda Leigh | 4.6 out of 5 stars |
| 19 | Beauty and the Billionaire: A Dirty Fairy Tale (Dirty Fairy Tales Book 1) | Lauren Landish | 4.3 out of 5 stars |
| 20 | The Old Girls' Network: A funny, feel-good read for 2020 | Judy Leigh | 4.1 out of 5 stars |
| 21 | Endeared (The Accidental Billionaires Book 5) | J. S. Scott | 4.7 out of 5 stars |
| 22 | That Boy: A Small Town, Friends-to-Lovers Romance (That Boy Series Book 1) | Jillian Dodd | 4.4 out of 5 stars |
| 23 | The Setup | Meghan Quinn | 4.5 out of 5 stars |
| 24 | That Wintry Feeling: A Novel (Debbie Macomber Classics) | Debbie Macomber | 4.2 out of 5 stars |
| 25 | Stealing Home (The Sweet Magnolias Book 1) | Sherryl Woods | 4.7 out of 5 stars |
| 26 | The Beekeeper's Promise | Fiona Valpy | 4.6 out of 5 stars |
| 27 | The Last Sister (Columbia River Book 1) | Kendra Elliot | 4.4 out of 5 stars |

ROMANCE 50

**Figure 15.** List of books for genre Romance

The data visualization through these figures not only demonstrates the structured approach taken in data analysis but also provides a practical reference for understanding the distribution of data across genres. These visualizations serve as key components in interpreting the results of the clustering and keyword extraction processes, which are further discussed in this chapter. By documenting these findings, the study demonstrates how the proposed methodology effectively streamlines data processing, improves data quality, and facilitates meaningful insights. The results serve as a basis for evaluating the performance of the system and its relevance to the objectives of the study, ultimately contributing to the development of a robust book recommendation model.

Figure 16 shows the data of reviews for fiction books that have been cleaned. This step will be repeated for all the reviews of the book. After all data is cleaned, we will choose only 200 reviews from each book because, due to hardware problems, it cannot run large data. The elbow method is one of the methods for selecting optimal K. The first step is calculating the squared error (SSE) sum for some K values. SSE is the sum of the square distance between the centroid and each member of the cluster. The plot will be several clusters K against an SSE graph. We will observe that SSE decreases and K increases as distortion is small. The position of the bend (knee) in the plot is usually considered to be an indication of the sufficient number of clusters.

| Review |
|---|
| witchcraft |
| decided coronapocalypse reread harry potter series likely looking read heard least bit series seen movie adult book easy read took two day casual reading finish could probably day anything else character fun engaging story move pretty fast book film adapt |
| may second third read first book book harry potter series still good read year book wa first published year old novel well written sympathetic character villainous antagonist house slytherin good plot consistent book well spread across book one story line k |
| heard name harry potter year reason another made attempt find fuss wa fast forward year later set year old triplet harry voldemort hagrid weasleys common name heard one night triplet dad house needed movie watch name harry potter came finally said |
| read book two reason year old granddaughter asked wanted able tell every hp fan asked read one never cared fantasy science fiction take review grain salt found pretty forgettable |
| perfect condition replace first one since wa stolen happier btw anyone read series stop everything buy book read series life change trust |
| amazing quite fit book though stupendous inspiring wonderful incredible quite describe either much better word put together read book please pick let anyone else negative opinion keep reading book ha much going friendship bravery courage love family m |
| probably person earth never read harry potter taking seriously amazon rating system star mean liked star mean loved liked harry potter sorcerer stone wa fun read creative humorous absorbing expected someone strongly prefers realism fantasy thought mi |
| late comer harry potter mania even interested reading book watching movie got free blu ray copy deathly hallows new blu ray player watched part rushed store buy first two movie see started viewing wanted read book found u scholastic version wa slightly |
| audio wa great cd poorly printed kind pale orange color assumed maybe screen printing wa completed could read information cd bow wa great shape wish could read cd know came first second third etc try cd order find first cd |
| admit wa late get board book started watching movie last book we released started getting book library year later bought paperback copy year read lead time finally got digital copy even though taken good care paperback afraid may fall apart latest endea |
| overall loved book wa best book ever read still think every day changed life pacing wa addicting put book wanted read every free moment plot plot wa perfect loved book wa best book read long time instant classic character felt real wa hook end chapter k |
| wa really confusing user stand maybe b c ten wa really weird half time wa thinking reading personally think get book like diary wimpy kid dork diary people age older like wizard magic think enjoy book |
| got book harry potter sorcerer stone old copy got lost move book book collection harry potter book reason fell love reading make believe world jk rowling thank writer created magical wizarding world inevitably give homey warm feeling end book want wiz |
| book great value afford hard cover book suggest invest beautiful grabbed husband wanted read hardcover parent house across country figure pa along local school done also say child dyslexic hardcover also joy read would frustrating |
| vividly described event event already since story lived mind people world want part world want read least base harry potter series understand friend family medium talking also recommened story anyone want learn widely accepted common practical englis |
| read book aloud family child period two week done many time book never wanted keep going going stop soon finished year old daughter read book ha become phenomenon hard see story appeal u matter age suspenseful part right amount keep u edge sea |
| love harry potter cd unabridged shortened version jim dale reading book voice fantastic love listen car driving really make awful commute go quickly get lost story much le stressed arrive destination |
| date fifth time read series love book glad daughter encouraged read first came librarian wa encouraged read could recommend student middle school classic adventure first rate much witchcraft good v evil win end would encourage kid age read |
| wa time started reading hp book wa expected magical |
| absolutely love harry potter series definitely one favorite series ever read first started reading really sure would like unfortunately thought wa nerdy wa school wa definitely huge mistake still good read adult really wish started reading came wanted kid able |
| wa wonderful revisiting story last time reread wa bought illustrated edition course started looking beautiful illustration next thing know wa lost world honestly never get old rowling writing marvelous pacing ability express character emotion know folk yet re |
| super mad book put together correctly example page end pick back page bought daughter th grade class super embarrassed editor pick seriously |
| love harry potter book film author j k rowling nail every way portrayal young wizard surround wondrously fantastic high fantasy go love love love story ability capture pure essence growing amongst many peer time story well harry potter story absolutely we |
| book movie great enjoy reason numerous mention book help happen board game comparing difference original book movie great conversation tidbit especially enjoy change book movie like |
| took long time choose reading first harry potter book primary hesitation wa child book sure wanted invest time we also concerned attention occult reference finished reading honestly say really sweet compelling story pleasant read rowling ha done fine job |
| year old recently discovered harry potter series grew conservative household harry potter wa banned wa child adult belief much liberal family heard one friend almost year senior talking much like harry potter series decided rent sorcerer stone movie liked r |

fiction

**Figure 16.** Data Cleaning

The data or information about the keyword, review genre, and title of each book are all visualized in Power BI. The dashboard consists of three parts. The first part of figure 17 is the visualization of the list of books that need to be focused on. The second part is the visualization of the result of clustering, keyword and details reviews for each genre, and the last part is a visualization of keywords for each book.

 ISSN 2723-6471



**Figure 17.** The dashboard shows a list of books

For the first part, the user can know the title of the book that will focused on and the number of totals of the book based on genre. In the right corner, the user can compare the rating between 4 stars and five stars. Slicer is used for rating purposes because slicers allow a user to sort and filter a structured report and view only the details they want. Below the rating is Genre, where the user can see the total of books for every genre, and the number of totals will show below and right on this figure. From this figure, we can see it has 162 books for all genres. The number will change when the user selects the genre that they want, as shown in figure 18 and figure 19 as examples.

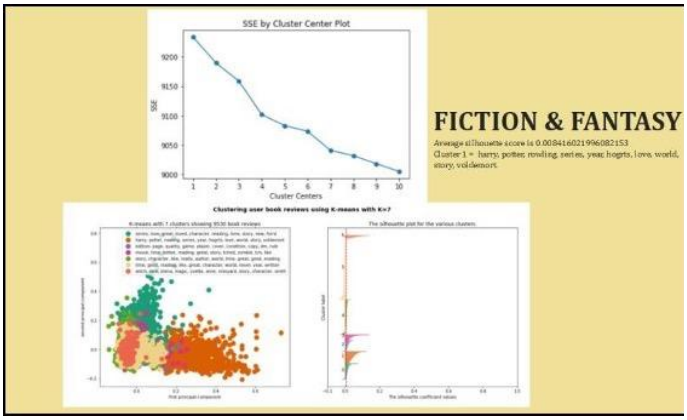



**Figure 18.** Result of clustering for genre Fiction and Fantasy


**Figure 19.** Keyword for a book on genre Fiction and Fantasy

This dashboard will help to identify which keywords are mostly mentioned in the review. By having the results can be taken into consideration in identifying and grouping similar data points in larger datasets without concern for the specific outcome.

## 5. Conclusions

This study successfully developed a recommender system for book reviews using the k-means clustering algorithm, addressing the challenge of navigating the overwhelming number of books available. By extracting and analyzing keywords from user-generated content on platforms such as Amazon.com, the system effectively streamlines the book recommendation process. The comprehensive methodology encompassed all key phases, including preliminary studies, data collection, data preprocessing, modeling, and visualization with Power BI.

The findings demonstrate that clustering techniques can effectively summarize and highlight trends in book reviews across various genres, making the recommendation process more precise. By analyzing and visualizing data based on commonly extracted keywords from reviews, the system provides tailored book suggestions that align with user preferences and interests. This enhances the user experience by simplifying decision-making and also offers valuable insights to publishers and authors, enabling a deeper understanding of reader preferences.

Future advancements could focus on integrating real-time data acquisition, improving the scalability of the system, and incorporating dynamic user input to refine the recommendations further. Enhancing these aspects would ensure the system remains robust and relevant in the ever-evolving digital book marketplace.

This study lays a strong foundation for developing more personalized and efficient digital book marketing and sales strategies, leveraging the power of artificial intelligence and machine learning in e-commerce. The recommender system demonstrates the potential to revolutionize the way books are marketed and consumed, ultimately benefiting both readers and the publishing industry.

## References

[1] H. Mondal, "The Book: In Scenario of 21st Century," *International Journal of Information and Computing Science*, vol. 2020, no. 7, pp. 7–16, 2020.

[2] C. Goldman, "This is your brain on Jane Austen, and Stanford researchers are taking notes," *Stanford News*, vol. 2012, no. 9, pp. 1–2, Sep. 2012.

[3] S. Y. Sari, F. R. Rahim, P. D. Sundari, and F. Aulia, "The importance of e-books in improving students' skills in physics learning in the 21st century: A literature review," *Journal of Physics: Conference Series*, vol. 2309, no. 1, pp. 1–10, 2022. doi: 10.1088/1742-6596/2309/1/012061.

[4] E. Gardiner and G. R. Musto, *The Oxford Companion to the Book*, Oxford University Press, vol. 2010, no. 9, pp. 1–20, Sep. 2010.

[5] A. Almjawel, S. Bayoumi, D. Alshehri, S. Alzahrani, and M. Alotaibi, "Sentiment Analysis and Visualization of Amazon Books' Reviews," in *International Conference on Computer Applications and Information Security (ICCAIS)*, vol. 2018, no. 1, pp. 1–6, 2018.

[6] R. Rani and R. Sahu, "Book Recommendation Using K-Mean Clustering and Collaborative Filtering," *International Journal of Engineering Sciences and Research Technology*, vol. 2017, no. 8, pp. 145–150, 2017.

[7] G. Pitsilis, X. Zhang, and W. Wang, "Clustering Recommenders in Collaborative," in *IFIP International Federation for Information Processing*, vol. 2011, no. 1, pp. 82–97, 2011.

[8] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "Recommendation systems: Principles, methods and practices," *Egyptian Informatics Journal*, vol. 2015, no. 1, pp. 261–273, 2015.

[9] D. C. Putri, J.-S. Leu, and P. Seda, "Design of an Unsupervised Machine Learning-Based Movie Recommender System," *Symmetry*, vol. 2020, no. 1, pp. 1–27, 2020.

[10] M. T. Himel, M. N. Uddin, M. A. Hossain, and Y. M. Jang, "Weight-based movie recommendation system using K-means algorithm," in *International Conference on Information and Communication Technology Convergence*, vol. 2017, no. 1, pp. 1302–1306, 2017.

[11] S. S. Manvi, N. Nalini, and B. Bhajantri, "Recommender system in ubiquitous commerce," in *International Conference on Electronics Computer Technology*, vol. 2011, no. 1, pp. 434–438, 2011.

[12] M. Garanayak, S. N. Mohanty, A. K. Jagadev, and S. Sahoo, "Recommender system using item-based collaborative filtering (CF) and K-means," *International Journal of Knowledge-based and Intelligent Engineering Systems*, vol. 2019, no. 1, pp. 93–101, 2019.

[13] J. Zeng, X. He, Y. Li, J. Wen, and W. Zhou, "A point of interest recommendation method using user similarity," *Web Intelligence*, vol. 2018, no. 1, pp. 105–112, 2018.

[14] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui, "GeoMF: Joint Geographical Modeling and Matrix Factorization for Point-of-Interest Recommendation," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 2014, no. 8, pp. 831–840, 2014.

[15] T. R. Kacchi and A. V. Deorankar, "Friend recommendation system based on lifestyles of users," in *IEEE International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics*, vol. 2016, no. 1, pp. 682–685, 2016.

[16] M. Sridevi, R. R. Rao, and M. V. Rao, "A survey on Recommender System," *International Journal of Computer Science and Information Security*, vol. 2016, no. 1, pp. 265–272, 2016.

 ISSN 2723-6471


[17] S. Surono, K. W. Goh, C. W. Onn, and F. Marestiani, "Developing an optimized recurrent neural network model for air quality prediction using K-means clustering and PCA dimension reduction," *International Journal of Innovative Research and Scientific Studies*, vol. 6, no. 2, pp. 330–343, 2023.

[18] H. Bhoir and K. Jayamalini, "Web Crawling on News Web Page using Different Frameworks," *International Journal of Computer Science and Engineering Information Technology*, vol. 2021, no. 7, pp. 1–5, 2021. doi: 10.32628/cseit2174120.

[19] A. Abodayeh, R. Hejazi, W. Najjar, L. Shihadeh, and R. Latif, "Web Scraping for Data Analytics: A BeautifulSoup Implementation," in *Proceedings of the Women in Data Science Conference, WiDS-PSU 2023*, vol. 2023, no. 1, pp. 1–6, 2023. doi: 10.1109/WiDS-PSU57071.2023.00025.

[20] N. A. Sultan and D. Abdullah, "Scraping Google Scholar Data Using Cloud Computing Techniques," in *Proceedings of the International Conference on Computing and Information Technology Management, ICCITM 2022*, vol. 2022, no. 1, pp. 1–7, 2022. doi: 10.1109/ICCITM56309.2022.10032044.

[21] S. Surono, K. W. Goh, C. W. Onn, and F. Marestiani, "Developing an optimized recurrent neural network model for air quality prediction using K-means clustering and PCA dimension reduction," *Int. J. Innov. Res. Sci. Stud.*, vol. 6, no. 2, pp. 330–343, Mar. 2023.