

Chapter 1: Introduction

1.1. Background and Significance of the Topic

The rapid advancement of technology has profoundly impacted consumer behaviour, particularly in their daily routines [1][2]. One of the most notable changes is the shift in trade transactions [3], with online shopping now surpassing offline shopping in demand and becoming an increasingly popular choice [4]. Transactions conducted over the internet, known as electronic commerce or e-commerce, have grown significantly in recent years [5][6]. This growth has been closely tied to technological advancements, each contributing to the expansion of e-commerce in different regions.

In Southeast Asia, Indonesia stands out as one of the fastest-growing digital markets. The rise of start-ups and fintech companies in Indonesia is a proof that Indonesia has a good environment for e-commerce [7]. Another contributing factor to the growth of e-commerce in Indonesia is the increasing number of internet users [8]. Indonesia is expected to account for 52 percent of the e-commerce market in Southeast Asia, with a market value projected to reach US\$150 billion in 2025 [7]. Moreover, by 2029, the number of e-commerce users in Indonesia is set to grow to 99.1 million, a 51.03% increase from 2024 [9].

With the rise of e-commerce, consumers are increasingly turning to online platforms to make purchasing decisions. However, navigating through various platforms to compare products can be time-consuming and inefficient [10]. Web scraping has become a popular solution to address this challenge, as it allows for the automated extraction of data from e-commerce websites [11][12][13] such as product prices, sales performance, and customers ratings [4][12][14]. This technology provides users with a more streamlined and efficient way to gather and compare product information in real-time, making it easier for them to make informed purchasing decisions [1][3][12].

Web scraping tools especially those utilizing Application Programming Interfaces (APIs), which enable programs to access website data directly in a structured format [15] allow users to gather information from various e-commerce platforms without the need to manually visit multiple websites, reducing time and effort while enhancing decision-making accuracy [11]. Moreover, the scraped data can be consolidated into a single spreadsheet file, allowing users to conveniently view and analyse information across multiple platforms [16].

This study addresses this need by developing a web application that utilizes Application Programming Interfaces (APIs) and machine learning to assist consumers in making informed online purchases. The proposed system allows users to input a product name, a target price, and a desired price range. It collects real-time product data from multiple Indonesian e-commerce platforms through available APIs and uses fuzzy string-matching to align similar product names [17][18]. The gathered data attributes such as price, number of sales, and customers ratings will be processed and normalized within range using a Min-Max Scaler [19][20]. The application generates weighted scores to recommend the most relevant products. This solution aims to empower consumers with a more efficient, accurate, and data-driven shopping experience.

1.2. Problem Statement

The rapid expansion and diversity of e-commerce platforms in Indonesia have created a fragmented shopping experience for consumers. With a multitude of online stores offering similar products, each with its own pricing, layout, and promotional strategies, consumers face significant challenges when comparing items across different sites. Key factors such as price, sales performance, and customers ratings are often presented inconsistently, leading to time-consuming and error-prone manual comparisons. As consumers frequently engage in price comparisons to secure the best deals, this repetitive behaviour significantly increases inefficiency and user frustration.

Although there are some APIs available for Indonesian e-commerce platforms, their implementation remains limited and fragmented. While these APIs can provide product data, they are not widely integrated into user-friendly tools designed to aggregate information from multiple platforms in one place. Most existing tools either focus on a narrow scope such as price comparison or fail to offer comprehensive, real-time updates on essential factors like customers ratings and sales performance. Additionally, these APIs often lack seamless integration within Indonesia's e-commerce platforms, resulting in missed opportunities to deliver an all-in-one solution for Indonesian consumers.

As a result, consumers are left to manually compare product listings, leading to inefficient and potentially flawed purchasing decisions. Without a unified solution that aggregates and presents key product data in real-time, shoppers face difficulty in making well-informed choices. A system that utilizes APIs to scrape data and uses machine learning to integrate factors like price, sales, and customers ratings would provide consumers with

accurate, data-driven recommendations, allowing them to make more informed and efficient purchasing decisions that offer the best value for their money.

1.3. Research Objectives

The primary objective of this project is to develop a web-based application that automates the collection and analysis of product data from Indonesian e-commerce platforms, specifically Tokopedia and Blibli. The system is designed to help users make informed purchasing decisions by comparing product prices, sales performance, and customers ratings across platforms.

This project focuses on creating a streamlined, easy-to-use interface for non-technical users, allowing them to input a product name, desired budget, tolerance range, and the number of search results. Based on this input, the system fetches relevant product data from multiple sources and ranks the products according to user preferences such as product name similarity, budget range, maximum desired results, and sorting criteria using filtering and matching algorithms.

1.4. Research Scope and Limitations

This study is limited to two selected Indonesian e-commerce platforms, Tokopedia and Blibli, which were chosen because the availability APIs for data extraction, allowing for more consistent and reliable data retrieval compared to traditional web scraping techniques. The system is designed to extract relevant product information from these platforms and provide recommendations based on user preferences.

While the system demonstrates the feasibility of cross-platform product comparison using API-based data collection and fuzzy string matching, several limitations must be acknowledged. Firstly, many popular e-commerce platforms in Indonesia including Shopee, Bukalapak, and Lazada implement strict anti-bot or anti-scraping mechanisms. These protections make it difficult, if not impossible, to programmatically access their data without violating their terms of service or encountering significant technical challenges. As a result, these platforms are currently excluded from the scope of the study.

The system also does not incorporate advanced features such as review sentiment analysis or image-based product similarity, which could further enhance the accuracy and relevance of product comparisons. The focus of this research remains on textual similarity and numerical attributes, leaving these more complex features for future development.

1.5. Research Methodology

This study follows a system development research methodology aimed at building an automated product comparison tool based on user preferences. The application helps users identify the best products based on factors like name similarity, price, sales performance and customers ratings across multiple e-commerce platforms.

- **Data Collection:** Product information is retrieved through public APIs from the selected e-commerce platforms and stored in DataFrame format for further processing.
- **System Designing:** A fuzzy string-matching algorithm is implemented to identify similar product names based on user input. The system also filters products by removing those outside the acceptable price range, calculated from the user's budget and tolerance percentage. Finally, products are sorted based on user-defined criteria, such as lowest price, highest sales, highest customers ratings, or a combination of these factors.
- **System Implementation:** A web-based prototype is built with a user-friendly interface to demonstrate the functionality.
- **Evaluation:** The system is tested and evaluated through user testing and scenario-based assessments to determine its accuracy, usability, and impact on decision-making.

1.6. Research Timeline

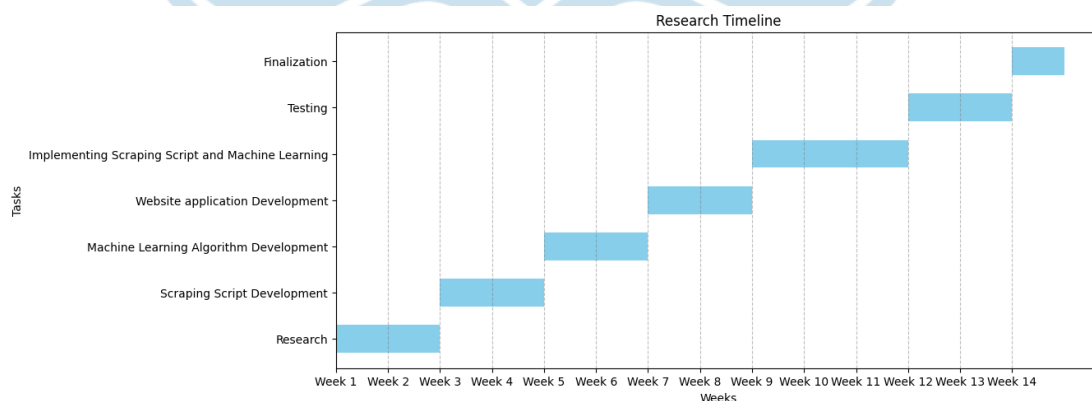


Figure 1 Research Timeline

1.7. Structure of the Thesis

This thesis is structured into six chapters, each contributing to the development and evaluation of a web application for scraping e-commerce data and recommending products using machine learning.

- Chapter 1: Introduction

This chapter provides the background and significance of the research, point out the problem statement, defines the research objectives, scope, and limitations, and outlines the methodology and timeline. It serves as the foundation for understanding the purpose and direction of the study.

- Chapter 2: System Analysis

This chapter presents an in-depth analysis of the technologies utilized in the project, analyse the feasibility of the system, specifies both functional and non-functional requirements, and compares various technical approaches considered during development.

- Chapter 3: Design

This chapter details the system's architectural and functional design, module structures, data schema, user interface layout, and scoring algorithm. It lays out the blueprint for how the system components integrate and interact.

- Chapter 4: Implementation

This chapter discusses the implementation of the application using Streamlit. It covers the development of the user interface, implementation of scraping e-commerce, data preprocessing functionalities, scoring system, and presentation of results to end users.

- Chapter 5: Testing

This chapter presents the results of module testing, evaluates the system's performance against the requirements, and identifies issues or bugs encountered during testing along with corresponding analyses.

- Chapter 6: Conclusion

This chapter summarizes the overall findings of the thesis, reflects on the accomplishments and limitations, and offers recommendations for future work and potential improvements.