CHAPTER I
INTRODUCTION

1.1 Background

A search system is a technology designed to assist users in finding relevant information based on a given input (query) [1]. The existence of an effective search system is crucial to expedite user access to accurate information [2]. In the academic context, an undergraduate thesis is a mandatory scholarly work that demands critical thinking in determining a research topic, reviewing existing literature, and formulating research methods [3]. Therefore, easy and precise access to prior theses is essential both as an academic reference and as a means of avoiding duplication of research topics.

At UPN "Veteran" Jawa Timur, thousands of theses have been documented in the form of physical archives as well as digital repositories. Although the digital repository has facilitated document access, the available search system remains limited, as it only returns relevant results when users enter a complete thesis title or author name. In practice, students frequently use brief queries or general terms related to their research topic, such as specific methods or fields of study. This condition causes the displayed search results to be less relevant and not reflective of user needs. These issues highlight the necessity of developing a thesis search system capable of comprehending the meaning of a query more effectively and producing relevant search results based on contextual and semantic similarity.

Prior research has developed a thesis search system for the UPN "Veteran" Jawa Timur repository using a content-based filtering method with TF-IDF weighting and weighted tree similarity, which has been demonstrated to be reasonably effective in generating keyword-based recommendations [1]. This method is capable of delivering relevant search results when the user's query matches words contained in the thesis title or abstract. However, this approach remains heavily dependent on lexical similarity, causing its effectiveness to diminish when users employ different terms that carry equivalent meaning.

Beyond keyword-based approaches, several other studies have developed document retrieval systems using different methods. Ranking approaches based on BM25 and the Vector Space Model (VSM) have proven effective in ordering

documents based on term frequency; however, they remain focused on lexical matching and are unable to capture semantic relationships between terms within the context of a sentence [4], [5]. Other studies have combined word embeddings with traditional term weighting to improve search relevance, yet this approach still relies on word-level representations and exhibits relatively high computational complexity [6]. Comparative studies on text representation have also shown that TF-IDF-based methods are inferior in understanding context compared to BERT-based models, particularly when queries employ synonyms or terminological variations [7].

The limitations of keyword-based approaches have driven the development of semantic search, an information retrieval approach that emphasizes the understanding of meaning (semantics) from both queries and documents, rather than merely matching words. This approach leverages embedding representations and transformer-based neural models to enable the system to retrieve relevant documents even in the presence of terminological discrepancies [8]. Semantic sentence embedding concepts such as SBERT enable meaning comparison between sentences using metrics such as cosine similarity [9]. Furthermore, query expansion techniques are employed to extend the initial query with relevant terms in order to bridge the linguistic gap between queries and documents. Modern contextual embedding-based approaches, such as CEQE, have been demonstrated to improve retrieval performance compared to static expansion methods [10]. For the Indonesian language, IndoBERT and its derivatives have proven more effective in capturing local contextual nuances and semantics compared to multilingual models, with improvements in academic search performance of 6-18% over Word2Vec [11].

Based on the aforementioned issues and limitations of prior research, this study proposes the use of IndoSentence-BERT (IndoSBERT) as the primary method for generating semantic representations (embeddings) of queries and documents. This model was selected for its capability to comprehend the meaning of sentences in the Indonesian language, enabling the search to extend beyond mere keyword matching to capture semantic similarity. In addition, this study applies a query expansion mechanism to broaden search keywords with other relevant terms,

thereby rendering the system more sensitive to overly short queries, abbreviations, synonyms, or terms not explicitly stated. For instance, when a user searches with the term “jaringan syaraf tiruan” (artificial neural network), the system will also associate it with the term “neural network”; or when the query contains “deep learning,” the search results will also encompass documents employing terms such as “cnn,” “xgboost,” or other related terms. The resulting embeddings are then matched using FAISS (Facebook AI Similarity Search) with cosine similarity computation, which ensures that the search remains both fast and accurate even as the volume of thesis data grows.

1.2 Problem Formulation

Based on the background described above, the problem formulation of this study can be stated as follows:

1. How can the thesis search system retrieve relevant theses based on meaning (semantics), even when there is a vocabulary discrepancy between the query and the thesis content?
2. How is the implementation of IndoSentenceBERT-based semantic search and query expansion applied within the thesis search system?
3. How does the semantic search system perform in handling the thesis collection in the UPN "Veteran" Jawa Timur repository, based on the evaluation results conducted?

1.3 Objectives

The objective of this research is to address the problem formulation presented above, namely to develop a semantic search-based thesis retrieval system capable of understanding semantic similarity between user queries and thesis titles or abstracts, even in the absence of explicit keyword matches, within the UPN "Veteran" Jawa Timur repository. The developed system implements IndoSBERT and query expansion, and supports the thesis search process on a large-scale data collection.

1.4 Benefits

The results of this research are expected to facilitate the process of searching for theses in the UPN repository by topic or research method through the implementation of a semantic search-based retrieval system capable of understanding semantic similarity between queries and thesis documents. This mechanism is realized through the utilization of semantic representations using the IndoSBERT method and query expansion applied to Indonesian-language academic documents.

1.5 Scope and Limitations

To ensure that the research remains within the determined scope, the following limitations are established:

1. The search system is focused on undergraduate thesis data from the UPN "Veteran" Jawa Timur repository.
2. The developed search system is focused solely on the core search functionality, with a simple user interface built using Streamlit to support system testing.
3. The query expansion method employed is word embedding-based, utilizing fine-tuned Word2Vec and a pre-trained FastText model that captures vector proximity based on the contextual similarity of word occurrences within the corpus.
4. Long-query testing is restricted to the domain of Informatics to maintain the quality and consistency of relevance assessment.