CHAPTER V
CONCLUSION AND RECOMENDATION

5.1 Conclusion

Based on the results of the research conducted, the semantic search-based thesis retrieval system using IndoSBERT has been demonstrated to be capable of retrieving relevant documents based on meaning (semantics), even in the presence of vocabulary discrepancies between the query and the documents. This capability is reflected in the high nDCG@15 values across all semantic models, with the highest achievement recorded in the SEM_FT scenario, which outperformed BM25 on long queries with an nDCG@15 of 0.8765. These results are further supported by query trials such as "loyalitas pelanggan" (customer loyalty), in which the semantic system successfully returned titles and abstracts with semantically relevant meaning, such as "kepuasan pelanggan" (customer satisfaction) and "brand trust." These findings indicate that the embedding-based approach is capable of representing semantic proximity between documents, enabling the system to identify topical relatedness even in the absence of explicit terminological overlap.

The implementation of IndoSBERT in this research was conducted through two approaches, namely the baseline model and the fine-tuned model, combined with a query expansion mechanism. Fine-tuning IndoSBERT on the UPN repository thesis domain was demonstrated to consistently improve ranking quality, as evidenced by the increase in nDCG@15 from 0.7653 (SEM_BASE) to 0.8265 (SEM_FT) on short queries. The application of query expansion made a selectively positive contribution, effectively improving retrieval coverage for short queries with minimal context, but proving ineffective for long queries that were already context-rich, as it introduced noise into the query vector representation.

In terms of performance, the system is capable of handling a collection of 15,326 thesis documents with an average retrieval time of approximately 0.5 seconds using FAISS. Ranking quality was maintained across all scenarios, with nDCG@15 ranging from 0.7653 to 0.8923, demonstrating that the developed semantic search system is feasible for deployment on large-scale document collections without sacrificing either efficiency or retrieval quality.

5.2 Recommendations

1. The semantic model can be further improved through the use of richer data and more comprehensive fine-tuning methods.

2. Further exploration of hybrid methods that directly combine lexical and semantic approaches (such as BM25 and embedding-based retrieval) is recommended, in order to simultaneously leverage the strengths of both approaches.

3. The query expansion mechanism within the semantic approach warrants further development, for instance through the use of context-based techniques or expansion weighting strategies, to avoid degrading precision.

4. System evaluation can be enhanced through the use of a gold standard dataset specifically designed to assess semantic similarity, such as query-document pairs based on synonyms or paraphrases. Through this approach, the ability of semantic search to handle vocabulary variation can be evaluated in a more measurable and objective manner, compared to the use of simple or keyword-based queries alone.