

Self-Reflection: NLP Final Project

Group Members: Ashwin Satish, Daniel Xiong, Roshan Chauhan

This project, focused on building a fact-verification system for the HoVer dataset, was a valuable learning experience. The most significant takeaways were understanding the trade-offs between traditional and neural retrieval methods and appreciating the engineering challenges of working with massive datasets. I gained a deeper appreciation for the practical differences between sparse retrieval (BM25) and dense retrieval (Sentence-BERT). While BM25 provided a strong baseline, our dense retrieval approach significantly improved recall after re-ranking, highlighting the power of semantic understanding. This project was also a lesson in data engineering. Handling the 30GB+ Wikipedia dump required careful pipeline design, and using tools like Elasticsearch and FAISS for indexing and similarity search was a crucial practical experience.

My primary contribution was leading the experimental dense retrieval portion of the project, as documented in the [dense_retrieval_faiss.ipynb](#) and [DenseRetr.ipynb](#) notebooks. I was responsible for setting up the FAISS index and implementing the re-ranking logic. This was a challenging task, but it led to a tangible improvement in our system's performance, boosting the Recall@10 by almost 10 percentage points and improving the final verification F1-score by 2 points.

If I were to do this project again, I would explore a hybrid retrieval approach from the start, combining the strengths of both sparse and dense retrieval in the initial retrieval step. I also would have liked to fine-tune the Sentence-BERT model on the HoVer dataset itself, which could have yielded even better results. This would have required more time but would have been a valuable next step.

I contributed about the same amount of work as my group members. We effectively divided the project into manageable parts: I focused on dense retrieval, while others handled BM25, claim verification, and project reporting was done equally by all, as each of us documented our work. We collaborated well and everyone contributed to the final report and presentation.

Given more time, I would have liked to further explore the experimental aspects of the project. While we met all core requirements, more time would have allowed for more experimentation with different pre-trained models, fine-tuning, and more advanced

hybrid retrieval techniques, which would have deepened our insights into fact verification.