Negation-Aware Retrieval for NevIR

GitHub: https://github.com/Hamad-Ayaz/NevIR-Project.git <- code for models.

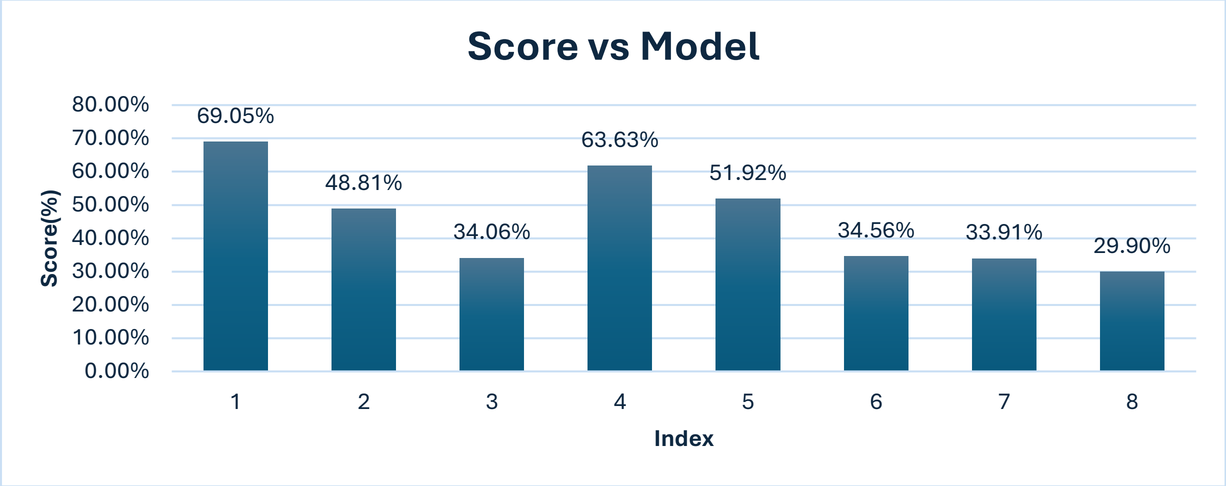

Fig.1: Score vs Model graph

Index	Type	Model Name	Params	Score
1	Cross-Encoder	QNLI-ELECTRA (Fine-Tuned)	110M	69.05%
2	Cross-Encoder	MS MARCO MiniLM	66M	48.81%
3	Cross-Encoder	QNLI-ELECTRA (Not Fine Tuned/Baseline)	110M	34.06%
4	Hybrid	TF-IDF + SBERT + FalconRW Negation	22M + 110M + 1B	63.63%
5	Hybrid	BM25 + SBERT-MPNet + RoBERTa	N/A + 110M + 355M	51.92%
6	Dense + LLM	BGE + DistilBERT Negation	180M + 66M	34.56%
7	Sparse + Prompt	TF-IDF + FalconRW Negation	1B	33.91%
8	Sparse	TF-IDF + Negation Flip	N/A	29.90%

Fig.2: Model Table

Negation is one of the most persistent challenges in information retrieval. The **NevIR** benchmark, introduced by Weller et al. (2024), was specifically designed to stress-test this issue by pairing near-identical documents that differ only in **negation polarity**—a setup that easily confuses models relying solely on lexical overlap. We use the **NevIR dataset** from this benchmark, which isolates negation as the key discriminative factor, making it an ideal testbed for evaluating negation-aware retrieval methods.

Traditional approaches such as **TF-IDF** fail almost entirely on this task, often scoring below **2%** (Pedregosa et al., 2011).

In this project, I experimented with a series of **sparse**, **dense**, **ensemble**, and **cross-encoder** models to address this issue. Each model is designed to test a specific hypothesis about negation handling in document ranking. Below is a detailed summary of each approach and its observed effectiveness on the **NevIR dataset**.

1. Sparse Models

TF-IDF Baseline:
TF-IDF computes similarity based on word frequency, but fails to detect semantic polarity. For instance, "Dogs are allowed" and "Dogs are not allowed" have high overlap, even though they convey opposite meanings. As a result, TF-IDF ranks both documents similarly, leading to **near-zero score** on NevIR.

TF-IDF + Negation Flip (29.90%):
To counter this, I introduced a **rule-based negation flip**, where queries containing words like "not", "never", or "no" trigger a reversal of the TF-IDF ranking. This simple fix alone boosts the score to **29.90%**, showing how even a naive negation-aware tweak can meaningfully improve retrieval.

2. Sparse + LLM Models

TF-IDF + FalconRW Negation (33.91%):
Instead of rules, I used prompting with a **1B parameter FalconRW LLM** to detect negation in queries. The prompt returned a binary “yes/no” for whether the query was negated. If yes, TF-IDF scores were flipped. This improved the score to **33.91%**, making it slightly better than the rule-based method and more robust to paraphrased negations (e.g., "I doubt this is true").

However, further prompting-based adjustments gave diminishing returns. Without fine-tuning the LLM, it became clear that **prompted negation detection had reached its performance ceiling** in this context.

3. Dense + LLM Models

BGE + DistilBERT Negation (34.56%):
This model combines a **dense retriever (BGE)** for semantic similarity and a **DistilBERT-based entailment classifier** to detect whether a query contradicts a given document. If negation was detected, the BGE similarity scores were reversed.

Despite using more powerful embeddings, this setup performed similarly to the LLM-prompted TF-IDF method, achieving **34.56%**. The likely cause is a lack of integration between the models—BGE ranks on similarity, while the classifier assesses entailment independently, leading to inconsistencies in voting.

4. Hybrid Models

TF-IDF + SBERT + FalconRW Neg (63.63%)
This ensemble model uses majority voting from three sources:

- **TF-IDF** for lexical overlap
- **SBERT** for semantic similarity
- **FalconRW** negation detection prompting

When evaluated independently, **TF-IDF + FalconRW** scores **33.91%**, and **SBERT** alone achieves **41.72%**. While both perform moderately well, combining them with simple majority voting results in a significantly improved **63.63%** score. Each component votes on which document better matches the query, and if two or more agree, that decision is selected.

This multi-view strategy captures distinct aspects of language—**term frequency**, **contextual meaning**, and **logical polarity**—making the ensemble more resilient to negation-induced ranking errors.

I also experimented with **weighted voting**, assigning greater influence to higher-performing models. However, this offered no consistent improvement over **majority voting**, suggesting the strength of the ensemble lies in its **complementary signals** rather than one dominant source.

BM25 + SBERT-MPNet + RoBERTa (51.92%):
I attempted to improve the hybrid by swapping in newer models:

- **BM25** in place of TF-IDF
- **SBERT-MPNet** in place of MiniLM SBERT
- **RoBERTa-large** MNLI in place of FalconRW

While these replacements are often considered stronger in general-purpose NLP tasks, this ensemble underperformed, achieving a **51.92%**.

This suggests that raw model power doesn't necessarily lead to better ensemble behavior. It's possible that the upgraded components were **less aligned in how they judged query relevance** or negation, resulting in **inconsistent decisions during voting**.

5. Cross-Encoders

MS MARCO MiniLM (48.81%):
Cross-encoders jointly process the query and document in the same input, enabling full attention across both. This pretrained **MS MARCO MiniLM** model, though not fine-tuned on NevIR, scored **48.81%**, outperforming many earlier models. Its ability to learn fine-grained contextual interactions makes it particularly suited for negation-sensitive tasks.

QNLI-ELECTRA (Not Fine-Tuned: 34.06% → Fine-Tuned: 69.05%):
To take full advantage of NevIR supervision, I fine-tuned **QNLI-ELECTRA**, a 110M parameter model originally trained on question entailment. Its baseline performance was modest (**34.06%**), but after fine-tuning with **pairwise margin ranking loss**, the model reached **69.05% score**—the highest of all tested models. The model directly learned to rank the correct document higher, even when negation was the only differentiating factor.

Conclusion

Fine-tuning a cross-encoder yielded the best overall performance. **QNLI-ELECTRA** (110M parameters), originally trained on question entailment, started with a modest baseline of **34.06%**. After fine-tuning on the NevIR dataset using pairwise margin ranking loss, it reached **69.05%**, the highest score across all tested models. The model directly learned to rank the correct document higher, even when **negation was the only differentiating factor**, showing the power of task-specific supervision.

The best trade-off between **simplicity**, **speed**, and **accuracy** was achieved by the **TF-IDF + SBERT + FalconRW** hybrid. Despite relying on relatively lightweight components, it reached **63.63%**, outperforming several more complex systems. This shows that combining **lexical**, **semantic**, and **logical** signals—even without training—can yield competitive performance.

In summary:

- **Prompting alone** is insufficient unless the model is fine-tuned.
- **Hybrid voting ensembles** show strong potential when carefully balanced.
- **Fine-tuned cross-encoders** provide the most accurate ranking signals but at a higher computational cost.

Future work could involve:

- Training hybrid models **end-to-end using multi-task loss**
- Using **learned voting weights** rather than fixed majority rules
- Exploring **reinforcement learning agents** for adaptive retrieval

Negation remains a difficult retrieval challenge—but with the right tools and supervision, it’s a **solvable** one.

Cited Works

Weller, O., Lawrie, D., & Van Durme, B. (2024). *NevIR: Negation in Neural Information Retrieval*. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2274–2287). Association for Computational Linguistics. https://aclanthology.org/2024.eacl-long.139/

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duch esnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.