A Hybrid Approach to Information Retrieval and Answer Generation for Regulatory Texts

Jhon Rayo
Universidad de los Andes
Bogotá, Colombia
j.rayom@uniandes.edu.co

Raúl de La Rosa
Universidad de los Andes
Bogotá, Colombia
c.delarosap@uniandes.edu.co

Mario Garrido
Universidad de los Andes
Bogotá, Colombia
m.garrido10@uniandes.edu.co

Abstract

Regulatory texts are inherently long and complex, presenting significant challenges for information retrieval systems in supporting regulatory officers with compliance tasks. This paper introduces a hybrid information retrieval system that combines lexical and semantic search techniques to extract relevant information from large regulatory corpora. The system integrates a fine-tuned sentence transformer model with the traditional BM25 algorithm to achieve both semantic precision and lexical coverage. To generate accurate and comprehensive responses, retrieved passages are synthesized using *Large Language Models* (LLMs) within a *Retrieval Augmented Generation* (RAG) framework. Experimental results demonstrate that the hybrid system significantly outperforms standalone lexical and semantic approaches, with notable improvements in Recall@10 and MAP@10. By openly sharing our fine-tuned model and methodology, we aim to advance the development of robust natural language processing tools for compliance-driven applications in regulatory domains.

1 Introduction

Information retrieval (IR) systems are concerned with efficiently querying large corpora to retrieve relevant results. Traditional systems, such as search engines, often depend on term-frequency statistical methods like *tf-idf*, which measures the importance of a term in a document relative to its frequency in the corpus (Melucci and Baeza-Yates, 2011). BM25 (Robertson et al., 1996), a well-established ranking function, builds on similar principles to provide a scalable and effective retrieval framework. However, such methods are inherently limited when addressing complex domains like regulatory texts, where the semantics often outweigh simple term matching.

Regulatory content is particularly challenging due to its specialized terminology and nuanced language. Synonyms, paraphrasing, and domain-specific jargon frequently obscure the relationship between queries and relevant documents, reducing the effectiveness of lexical retrieval methods.

Semantic search addresses these limitations by using dense vector-based retrieval where we encode documents and queries as vectors, also known as *embeddings*, capturing the semantic meaning of the text in a condensed high-dimensional space (Karpukhin et al., 2020). This approach enables the system to measure similarity based on meaning rather than exact word matches, grouping related content together even with different terminology.

Recent advances in pre-trained language models, like BERT (Devlin et al., 2018), have introduced high-quality contextual *embeddings* for words, sentences, and paragraphs which can be leveraged in semantic search applications.

Despite these advances, building an effective IR system for regulatory texts poses unique challenges. Pre-trained language models are typically trained on general-purpose datasets and may lack the domain-specific knowledge required for accurate retrieval in specialized fields. Fortunately, various methods for transfer learning have demonstrated that these base models can be fine-tuned to close this gap (Houlsby et al., 2019).

In this paper, we present a hybrid information retrieval system that integrates both lexical and semantic approaches to address the limitations of traditional IR in the regulatory domain. Our method combines BM25 for lexical retrieval with a fine-tuned Sentence Transformer model (Reimers and Gurevych, 2019) to improve semantic matching. Additionally, we implement a Retrieval Augmented Generation (RAG) system (Lewis et al., 2021) that leverages the hybrid retriever to provide comprehensive and accurate answers to user queries using a Large Language Model (LLM).

Through extensive experiments, we demonstrate that the hybrid retriever achieves superior perfor-

Proceedings of the 31st International Conference on Computational Linguistics, pages 31–35
January 19–24, 2025.

mance compared to standalone lexical or semantic systems, as evidenced by improvements in Recall@10 and MAP@10. Furthermore, the RAG system effectively synthesizes retrieved content, delivering detailed responses that address the compliance requirements of regulatory questions. Our contributions aim to advance regulatory information retrieval and lay the foundation for more effective question-answering systems in specialized domains.

2 Regulatory Information Retrieval

The development of an effective information retrieval (IR) system for regulatory content requires addressing the unique challenges of compliance-related queries. These systems must return a set of ranked passages from the corpus that accurately address the compliance aspects of a given question. Previous work by Gokhan et al. (2024) utilized BM25, a widely-used algorithm that ranks results based on query term frequency and other statistical features. While BM25 is effective for lexical retrieval, it struggles to capture semantic relationships, particularly in regulatory domains where terminology often varies for the same concepts.

Our approach enhances BM25 by integrating a text embedding model, enabling semantic matching. This hybrid system identifies semantically relevant content that BM25 alone might overlook, offering a significant advantage in handling the complexities of regulatory language.

2.1 Dataset

The dataset used for this study, ***ObliQA***, consists of 27,869 regulatory questions extracted from 40 documents provided by Abu Dhabi Global Markets. This regulatory authority oversees financial services within the European Economic Area, making the dataset highly relevant for compliance-related tasks (Gokhan et al., 2024).

The dataset is divided into three subsets: training (22,295 questions), testing (2,786 questions), and validation (2,788 questions). Each question is paired with one or more passages that contain the relevant information needed to answer it. The data is stored in JSON format, where each entry includes the question, associated passages, and their metadata. An example is shown below.

```
{
  "QuestionID":
      ↪ "a10724b5-ad0e-4b69-8b5e-792aef214f86",
  "Question": "What are the two specific
      ↪ conditions related to the maturity of
      ↪ a financial instrument that would
      ↪ trigger a disclosure requirement?",
  "Passages": [
    {
      "DocumentID": 11,
      "PassageID": "7.3.4",
      "Passage": "Events that trigger a
          ↪ disclosure. For the purposes of
          ↪ Rules 7.3.2 and 7.3.3, a Person is
          ↪ taken to hold Financial ..."
    }
  ],
  "Group": 1
}
```

2.2 Model Fine-tuning

We fine-tuned the *BAAI/bge-small-en-v1.5* (Xiao et al., 2023), a BERT-based model trained on general-purpose data. The fine-tuning process employed a loss function designed to maximize the similarity between questions and their associated passages. The architecture comprises a word embedding layer followed by pooling and normalization layers. To better capture semantic nuances in regulatory texts, we increased the embedding dimension from 384 to 512.

Training was conducted on an NVIDIA A40 GPU with 24GB of memory using the *SentenceTransformer* library (Reimers and Gurevych, 2019). The model was trained over 10 epochs with a batch size of 64, using a learning rate of $2x10^{-4}$ to preserve the model's general-purpose knowledge while fine-tuning it for the domain. The *MultipleNegativesRankingLoss* (Reimers and Gurevych, 2023) loss function was employed, assuming all unpaired examples in the batch as negatives, which is particularly suited for scenarios with positive pairs only.

Performance evaluation was conducted using the *InformationRetrievalEvaluator* (Reimers and Gurevych, 2021) to compute metrics such as Recall@10, Precision@10, and MAP@10 during training. To further optimize the process, we employed warmup steps to gradually increase the learning rate, and Automatic Mixed Precision (AMP) (Zhao et al., 2021) to reduce memory usage and enhance training speed.

Table 1 summarizes the results, showing a significant performance improvement of the fine-tuned model over the base model in the regulatory domain. The fine-tuned model has been made avail-

Model / Dataset	Recall@10	MAP@10
Base Model / Validation	0.7135	0.5462
Base Model / Testing	0.7017	0.5357
Custom Model / Validation	0.8158	0.6315
Custom Model / Testing	**0.8111**	**0.6261**

Table 1: Performance comparison between the base model and the fine-tuned model.

able on Hugging Face Hub, alongside the complete implementation in our GitHub repository.

2.3 Information Retrieval

To enhance retrieval performance, we developed a data processing pipeline with the following steps:

1. **Expand contractions**: Convert contractions (e.g., *don't* to *do not*) for consistency.
2. **Normalization**: Lowercase text and remove non-alphanumeric characters using regular expressions.
3. **Space removal**: Eliminate redundant spaces for uniformity.
4. **Preserve legal format**: Retain special characters critical for legal documents.
5. **Stopwords**: Remove common words using *nltk* and *scikit-learn* sets.
6. **Stemming**: Apply the *Snowball Stemmer* (Porter, 2001) to reduce words to their root forms.
7. **Tokenization**: Generate unigrams and bigrams to capture both individual terms and word combinations.

Using this pipeline, we implemented three retrieval approaches:

1. BM25 (Baseline): Configured with $k = 1.5$ and $b = 0.75$.
2. Semantic Retriever: Leveraged the fine-tuned model for semantic matches only.
3. Hybrid System: Combined BM25 and the fine-tuned model, computing an aggregated score using Equation 1:

$$\text{Score} = \alpha \cdot \text{Semantic Score} + (1 - \alpha) \cdot \text{Lexical Score} \quad (1)$$

Model	Recall@10	MAP@10	Recall@20	MAP@20
BM25 (Baseline)	0.7611	0.6237	0.8022	0.6274
BM25 (Custom)	0.7791	0.6415	0.8204	0.6453
Semantic system	0.8103	0.6286	0.8622	0.6334
Hybrid system	**0.8333**	**0.7016**	**0.8704**	**0.7053**

Table 2: Performance comparison between information retrieval systems.

We empirically set $\alpha = 0.65$ to give slightly higher weight to semantic matching while maintaining meaningful contribution from lexical search. This normalization step ensures that neither approach dominates the final ranking purely due to differences in score distributions.

Table 2 compares the performance of these approaches. The hybrid system demonstrates the highest effectiveness, combining the strengths of lexical and semantic retrieval methods.

3 Answer Generation

Retrieval Augmented Generation (RAG) is a cutting-edge technique that enhances *Large Language Models* (LLMs) by integrating external retrieval capabilities, enabling them to generate responses based on information they were not explicitly trained on (Lewis et al., 2021). This approach has emerged as a powerful tool in open-domain question-answering applications, combining retrieval-based and generation-based methods to improve answer relevance and quality (Siriwardhana et al., 2023).

In our system, RAG is used to answer regulatory questions by leveraging the hybrid information retrieval system described earlier. The retrieved passages provide the contextual foundation for generating answers that address compliance-related aspects comprehensively and accurately.

Given a regulatory question, similar to the approach followed in (Gokhan et al., 2024), the system retrieves up to 10 relevant passages from the corpus. To ensure high-quality input for the answer generation process, only passages with a relevance score of at least 0.72 are considered. Additionally, passage processing is terminated when the relevance score drops by more than 0.1 from the previous passage, maintaining the relevance and coherence of the input data.

These selected passages are fed into an LLM to synthesize a concise and coherent answer. For this task, we experimented with three different models: *GPT 3.5 Turbo* and *GPT-4o Mini* through Azure OpenAI

batch deployment, and *Llama 3.1* using Groq's API. When evaluated on our test dataset, *GPT 3.5 Turbo* achieved the highest RePASs score of 0.57, significantly outperforming both *GPT-4o Mini* (0.44) and *Llama 3.1* (0.37), leading to its selection as our primary model. We designed the system prompt to guide response generation in the regulatory domain, emphasizing accuracy, completeness, and alignment with the provided passages. The prompt reads:

> "*As a regulatory compliance assistant. Provide a **complete**, **coherent**, and **correct** response to the given question by synthesizing the information from the provided passages. Your answer should **fully integrate all relevant obligations, practices, and insights**, and directly address the question. The passages are presented in order of relevance, so **prioritize the information accordingly** and ensure consistency in your response, avoiding any contradictions. Additionally, reference **specific regulations and key compliance requirements** outlined in the regulatory content to support your answer. **Do not use any extraneous or external knowledge** outside of the provided passages when crafting your response.*"

We selected the top 3 answers with the highest RePASs scores to enhance the prompt using few-shot techniques, aiming to improve its performance. Below is a demonstration of how we used this prompting method.

> " *Question: What percentage of the Insurer's Net Written Premium is used to determine the non-proportional reinsurance element? Passage: The non proportional reinsurance element is calculated as of the Insurer's Net Written Premium Your response should read: The non-proportional reinsurance element is determined by calculating 52 percent of the Insurer's Net Written Premium.*"

Regulatory Passage Answer Stability Score (RePASs), introduced by Gokhan et al. (2024) assesses the stability and accuracy of generated answers across three key dimensions:

1. Entailment Score (E_s): Measures the extent to which each sentence in the generated answer is supported by sentences in the retrieved passages.

2. Contradiction Score (C_s): Evaluates whether any sentence in the generated answer contradicts the information in the retrieved passages.

3. Obligation Coverage Score (OC_s): Checks if the generated answer covers all obligations present in the retrieved passages.

System	Es	Cs	OCs	RePASs
Baseline	0.78	0.24	0.20	0.58
Hybrid retriever + GPT-4o Mini	0.38	0.23	0.17	0.44
Hybrid retriever + Llama 3.1	0.34	0.45	0.22	0.37
Hybrid retriever + GPT 3.5 Turbo	**0.58**	**0.21**	**0.33**	**0.57**

Table 3: Performance comparison of answer generation systems using RePASs metrics.

The composite RePASs score is derived from these metrics, offering a holistic measure of the system's answer quality. Table 3 summarizes the evaluation results, comparing our approach to the baseline.
Table 3 shows that while our system achieves moderate improvements in obligation coverage (OC_s) and slightly better contradiction handling (C_s), its entailment score (E_s) reveals areas for further optimization. The hybrid retrieval system enhances answer relevance by incorporating semantic and lexical matches, but the synthesis process using *GPT 3.5 Turbo* shows reduced performance in capturing the degree to which generated answers are supported by the retrieved passages, as evidenced by the lower entailment score.

4 Conclusion

This work tackles the significant challenges of retrieving and synthesizing information from complex regulatory texts by demonstrating the effectiveness of hybrid approaches that integrate lexical and semantic retrieval methods. Our results show the importance of combining classical algorithms, such as BM25, with embedding-based models to address the nuanced language and diverse terminologies inherent in regulatory domains. The hybrid system consistently outperforms standalone lexical or semantic approaches, achieving notable improvements in metrics like Recall@10 and MAP@10.
We further demonstrate the potential of LLMs to synthesize concise and comprehensive answers. These models effectively utilize the structured information retrieved by the hybrid system to address regulatory queries with improved coherence and relevance. However, the evaluation using RePASs reveals opportunities for refinement, particularly in improving entailment metrics.
Future directions include fine-tuning LLMs on domain-specific corpora to enhance alignment with regulatory contexts, optimizing retrieval thresholds for better semantic coverage, and exploring advanced scoring mechanisms to balance precision and recall.

Acknowledgments

This work was supported by the NLP Group at Universidad de los Andes. We thank Abu Dhabi Global Markets for providing access to their regulatory documents. Special thanks to our dedicated professor Rubén Francisco Manrique.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Tuba Gokhan, Kexin Wang, Iryna Gurevych, and Ted Briscoe. 2024. Regnlp in action: Facilitating compliance through automated information retrieval and answer generation. *Preprint*, arXiv:2409.05677.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *Preprint*, arXiv:1902.00751.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Preprint*, arXiv:2004.04906.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Massimo. Melucci and Ricardo. Baeza-Yates. 2011. *Advanced Topics in Information Retrieval*, 1st ed. 2011. edition. The Information Retrieval Series, 33. Springer Berlin Heidelberg, Berlin, Heidelberg.

Martin F. Porter. 2001. Snowball: A language for stemming algorithms.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2021. Information retrieval evaluator. `https://sbert.net/docs/package_reference/evaluation.html#sentence_transformers.evaluation.InformationRetrievalEvaluator`.

Nils Reimers and Iryna Gurevych. 2023. Sentence transformers documentation: Losses. `https://sbert.net/docs/package_reference/losses.html#multiplenegativessymmetricrankingloss`.

Stephen E Robertson, Steve Walker, MM Beaulieu, Mike Gatford, and Alison Payne. 1996. Okapi at trec-4. *Nist Special Publication Sp*, pages 73–96.

S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

C. Zhao, Ting Hua, Y. Shen, L. Qian, and H. Jin. 2021. Automatic mixed-precision quantization search of bert. *Preprint*, arXiv:2112.14938.