Optimizing the Dutch Open Government Act Information Retrieval

by

Yao Hua Ju

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday November 22, 2024 at 01:30 PM.

Student number:	4856708	
Project duration:	Febuary 19, 2024 – November 22, 2024	
Thesis committee:	Prof. dr. ir. C. S. S. Liem	TU Delft, Chair, Thesis Advisor
	Prof. dr. ir. M. S. Pera	TU Delft, Core Member
External members:	V. Gevers	SSC-ICT, Innovation Manager
	R. Fiedler	SSC-ICT, Innovation Manager

This thesis is conducted in collaboration with SSC-ICT.

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Preface

I remember my first few weeks of starting the BSc in Computer Science & Engineering here at the Delft University of Technology very well. Coming in with no prior coding experience, I noticed how many of my fellow students had a significant head start. I struggled hard, and I still remember that I did not understand how to print a triangle of stars in Java (now I know its by using a double loop). Fast forward to today, and here I am, on the brink of graduating with my MSc in Computer Science. It has been a long ride, filled with countless late nights at the UB together with countless hours of lectures. It was not an easy journey, but looking back, I am proud of how far I have come and how much I have grown, both academically and personally.

First and foremost, I would like to thank my Thesis Advisor, Ms. Liem. Thank you for guiding me over the past nine months to what I can now call my final academic paper. There's a short funny story that I have never shared with you before. Five years ago, during the Signal Processing course, you gave a lecture somewhere halfway through the quarter. This single lecture stood out to me so much that I approached you after class, to ask if you would be giving more lectures. I was only disappointed to hear that that would not be the case. Then, I would be fortunate enough to have you as my supervisor for my master's thesis. I'm truly grateful for your guidance and support throughout this process.

I would also like to thank Ms. Pera for taking the time to be part of my thesis committee. I very much enjoyed the Information Retrieval course that you co-taught, and your enthusiasm for teaching left a strong impression on me. In fact, I found the subject so engaging that it inspired me to focus my thesis primarily on Information Retrieval.

Thank you, Victor and Ramon, for being amazing supervisors at SSC-ICT. Thank you Jitse, Martha, Daan, Lara, Ahmed, Mark, and the entire SSC-ICT team, for being amazing colleagues. And thank you to everyone who has made time to support me during my defense, whether in person or in spirit, thank you for being there for me. Without all of you, I would not be here today.

Yao Hua (Nicky) Ju
Delft, November 2024

Abstract

In 2022, the Dutch Open Government Act (Wet open overheid, Woo) has required that government institutions share requested documents with citizens, thereby enhancing government transparency and public access to information. However, current document retrieval processes often struggle to meet the legal requirements of the Woo, as they frequently fail to respond to requests within the legally mandated time frame due to the lengthy retrieval process.

This study addresses the technical challenges of optimizing information retrieval systems in the context of the Woo, by focusing primarily on document precision and recall. By critically analyzing existing workflows, we identify key inefficiencies and propose enhancements. Our research includes a comparative evaluation of dense and sparse retrieval methods to assess their effectiveness in this domain. Additionally, we explore different preprocessing techniques, investigating their impact on retrieval performance on both sparse and dense retrieval systems, to determine the optimal approach for handling noisy, unstructured government data.

Our results show that these changes in retrieval methods can significantly improve retrieval accuracy and reduce response times. BM25 in particular, shows strong performance, effectively handling the noisy data often present in government documents, highlighting its suitability for this context. These insights provide insights for government institutions to improve and streamline their information retrieval workflows, and reduce delays of the Woo requests.

Contents

List of Figures

List of Tables

1

Introduction

In recent developments within the context of government transparency, the Dutch government is now required to proactively share information under the Open Government Act (in Dutch: Wet open overheid, Woo)[1]. Enacted on the 1st of May 2022, this act states that specific sections of government data must be made accessible to the public. Details on this act can be found in article 5.1[2] and article 5.2[3] of the Woo. Examples of data that need to be made public are reports, e-mails, documents, or even WhatsApp messages and other text messages [11].

In the case that government data is not yet publicly available, people are also allowed to request this specific data, also referred as a Woo request. The request can be submitted to any government organization and can contain information such as how the government acts, or why and how a decision was made [40]. The government is, by law, required to respond to the request within 4 weeks. However, in the case that a request is too complex, is too long, or has a different valid reason, the government is allowed to extend this period with a maximum of 2 weeks [59]. Unfortunately, government agencies often fail to respond within this time duration. Looking at statistics shared by the Dutch government in 2022 [41], 75% of the Woo requests that were being processed by Dutch ministries were not answered within the legal deadlines [43]. From these 75%, about 142 (about 20% of the total) already took longer than a year. The longest request still pending at that time was 1568 days. Due to lengthy resolve times for requests, ministries can experience fines exceeding 1.8 million Euros in just a few months [42]. Director Wiemers of the Open State Foundation believes that an open government fosters citizen trust[1], but with such figures, regaining trust becomes a daunting challenge. This inefficiency hinders access to information and undermines the government's goals of accountability and transparency, all while resulting in significant financial penalties.

A recent article published by the government [21] also states that the cabinet is taking extra measurements to improve the process of Woo requests. This clearly shows that the government is aware of the issues and is actively working to improve processes to ensure that they are following the law.

1.1. Problem Statement

The long processing time for Woo requests cannot be attributed to a single factor; instead, it results from a combination of multiple issues. Director Wiemers points out that a significant contributing factor is the prevailing mentality within departments of considering the handling of

[1] https://www.rijksoverheid.nl/onderwerpen/wet-open-overheid-woo
[2] https://wetten.overheid.nl/BWBR0045754/2023-04-01/#Hoofdstuk5_Artikel5.1
[3] https://wetten.overheid.nl/BWBR0045754/2023-04-01/#Hoofdstuk5_Artikel5.2

these requests as a chore [37]: currently, the public disclosure of documents is not treated as a priority but rather as something government officials attend to only after completing their regular duties. Another reason for the long processing time is the large amount of manual work that still has to be done; think about the manual removal of personal information in the given documents, or the physical locating and scanning of documents that are not already in computer-readable format. This labor-intensive process can significantly delay the availability of information, as it requires thorough attention to detail and consumes a considerable amount of time and resources. However, discoverability is considered one of the biggest problems. Within the entire process of digital collaboration, there is most dissatisfaction about finding documents, as many government officials struggle with locating information effectively in their Document Management Systems [6]. Almost all government officials (96%) indicate that clear search results play a (very) important role in organizing and maintaining information management, highlighting the need for improvements in discoverability and access to information within government systems.

This concern over discoverability emphasizes the need for effective information retrieval systems within government operations. Information retrieval is a field of research focused on the effective and efficient retrieval of relevant information from large-scale corpora [4, 60]. Effective information retrieval is not just about finding data, but also about ensuring that the retrieved information is relevant and timely. This process is vital for various applications including information extraction and question answering [36]. In the context of government operations, efficient information retrieval systems can significantly enhance the transparency and accountability required by the Woo. However, the current challenges in the discoverability make it hard to meet these objectives.

Therefore, improving findability within the governmental information retrieval systems is critical. Enhancements in this area could streamline the Woo request process, reduce delays, and ultimately strengthen public trust in government transparency initiatives. In this paper, we will specifically explore the issue of findability in the context of information retrieval with government documents, examining its impacts and proposing solutions to improve it. A thorough explanation of the current process of Woo requests within ministries can be found in Section 2.1.

1.2. Research Questions

Given the need for reliable access to government information, it is clear that current workflows fall short in meeting these demands properly. Improving this is essential to improving transparency and fostering an open government. This thesis aims to explore the issues impacting Woo requests, focusing on information retrieval within Dutch ministries. By analyzing the current state of information retrieval methods, investigating potential technological improvements, and evaluating the speed and performance of various retrieval systems, we seek to provide a comprehensive view of how to improve the information retrieval for handling of Woo requests.

This leads us to the following research questions:

- **RQ1: How effective are current state-of-the-art information retrieval methods in finding Woo requested information within Dutch Ministries?**
- **RQ2: What technological improvements can Dutch Ministries implement to enhance the accuracy and efficiency of document retrieval for Woo requests?**
- **RQ3: How quickly can information from databases be created and retrieved?**

All code used in this research, including scripts for data preprocessing, model implementation, and analysis, is publicly available on GitHub to ensure transparency and allow for reproducibility of the results (accessible at https://github.com/SSC-ICT-Innovatie/LearningLion-WOO). More detailed instructions on how to run the code can be found in the repositorys README.

1.3. Motivation

SSC-ICT is one of the largest ICT service provider for the Dutch government. In SSC-ICT, the innovation team has been tasked with maintaining a forward-looking approach to their projects. In alignment with this vision, they have started the development of LearningLion [4], which is a project on the use of generative AI to improve services provided by SSC-ICT. The project aims to streamline and optimize internal processes, providing better support to employees and improving overall organizational efficiency.

The focus of SSC-ICT within this initiative, is on the generative Large Language Models (LLM), through the use of Retrieval-Augmented Generation (RAG).

This approach integrates the retrieval of relevant documents with generative AI to produce accurate and contextually appropriate responses, thereby improving decision-making and operational workflows. While RAG offers an innovative approach, there remains an awareness that RAG will not necessarily solve all problems in retrieval. It is approached as one potential improvement within a broader framework of enhancements, with careful consideration of both its benefits and possible limitations. The current flow design that SSC-ICT employs for this purpose is illustrated in Figure 1.1.

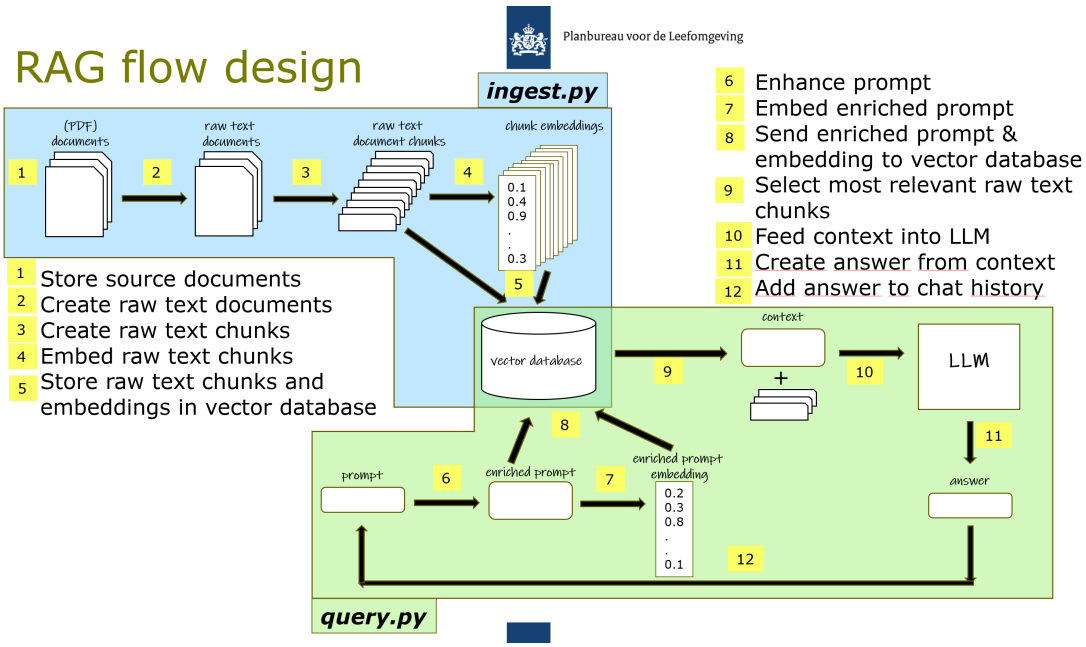

Figure 1.1: Current RAG flow design, used by SSC-ICT.

For this thesis, the primary focus will be on enhancing the retrieval process of relevant documents within this RAG framework. Specifically, the research and development efforts will target improvements in steps 1 through 8 as depicted in Figure 1.1. It is important to note that RAG systems typically use embedding-based (dense) retrieval methods. However, in this paper, we will broaden the scope by considering not only these dense retrieval systems but also sparse retrieval systems, allowing for a more comprehensive analysis of retrieval approaches. By refining these initial stages, the goal is to ensure that the system retrieves the most accurate documents, thereby laying a stronger foundation for subsequent generative AI processes (not limited to just Woo) and ultimately improving the quality and effectiveness of SSC-ICT's services.

[4] `https://learninglion.nl/`

2

Preliminaries

2.1. Current Workflow

The government and its various ministries currently manage their documents primarily using Digidoc, a Document Management System (DMS) designed to support the digital operations and archiving needs of employees. While Digidoc is the primary tool, as can be seen in the top section of Figure 2.1 other document management systems may also be used to meet specific requirements and depending on where data is already stored. Digidoc enables government officials to upload documents to a cloud-based repository, allowing for easy retrieval for activities such as information requests. In addition to the files, users are required to complete various fields with metadata to enhance the organization and findability of these documents. Despite these features, the implementation often suffers from inconsistent file uploading practices. This inconsistency not only hinders the efficiency of document retrieval but also affects the overall integrity of the document management system. Even though Digidoc is considered the main tool for governmental institutes, their accessibility is still considered 'D Tier'[1]. This means that the website does not comply to legal obligations but there is also no plans to currently increase this accessibility. Aside from that, the former Dutch Secretary of Digitalization Van Huffelen has also mentioned that the government needs to make more information accessible in one central place, rather than at multiple locations [37].

Together with Digidoc, the government uses Search & Find (in Dutch: Zoek & Vind), an Enterprise Search Application [67] to look through their database. An image of the home page of Search & Find can be found in figure 2.1. This tool enables comprehensive search across multiple databases, allowing users to locate specific information based on their queries. It can not only look inside of the DMSes, but also across network drives, websites, email inboxes, and other resources. With a single query, users can access a wide array of data sources, significantly enhancing the efficiency and effectiveness of governmental digital searches. The algorithm utilized for these search operations is based on boolean search. While boolean search is a simple search algorithm, it naturally comes with advantages and disadvantages, especially in the setting of governmental digital searches.

One of the primary advantages of boolean search is the precision and control it offers. By combining keywords with operators like AND and OR, users can refine their queries to very specific conditions. This level of precision is particularly beneficial in a governmental setting where the recall of information retrieval is essential. As it is important that as much relevant data as possible is getting retrieved to provide the completeness for the user. Essentially, boolean search provides flexibility in query formulation, allowing users to create complex queries that narrow down results

[1] `https://dashboard.digitoegankelijk.nl/organisaties/678/websites-apps/4004`

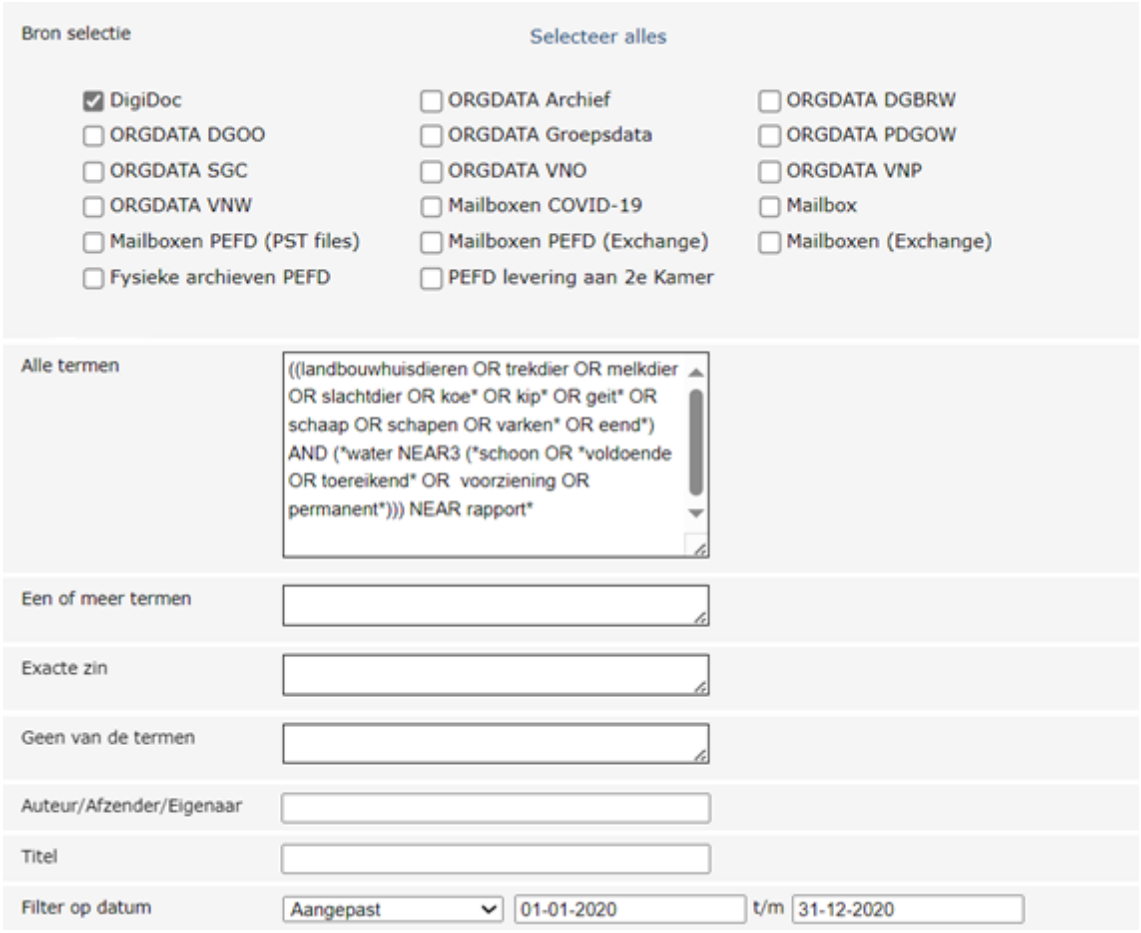

Figure 2.1: A screenshot of the home page of Search & Find.

to the most relevant documents. This flexibility is crucial when dealing with vast amounts of data spread across multiple databases, network drives, websites, and email inboxes. The efficiency of boolean search in handling large datasets quickly also stands out, significantly reducing the time needed to locate critical information and thus improving productivity and decision-making processes.

Despite its precision and flexibility, boolean search can be unwieldy and demanding for users who may not be familiar with the exact syntax required to construct effective queries. This rigidity often results in a steep learning curve and may lead to frustration or inefficient searches if the query is not perfectly formulated. Although document ranking is essential for many information retrieval related tasks [65], boolean search lacks a ranking mechanism, which means it does not prioritize results by relevance, leading to significant user effort in sifting through potentially large datasets to find the most relevant information. This absence of ranking can be particularly problematic in complex informational landscapes where time is critical. Now, Woo coordinators have to go through dozens of files, before finding one that is potentially relevant. Lastly, Boolean search overlooks the nuances of language such as synonyms, stemming, and semantics, which can cause it to miss critical documents that do not match the query exactly but are still relevant. This limitation reduces the overall effectiveness of the search tool in dynamic and diverse information environments like those of governmental operations.

To make the process as smooth as possible, there is a workflow currently used by Woo coordi-

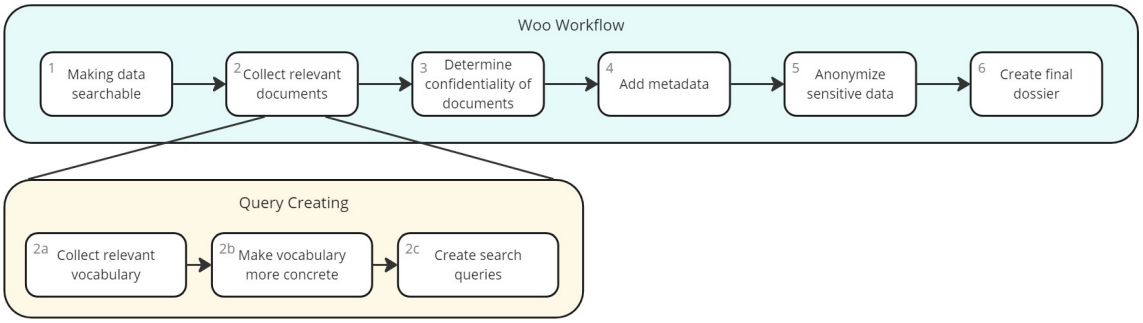

Figure 2.2: Workflow used by Woo coordinators.

nators, which can be seen in figure 2.2. These steps are all manually executed, without help from the computer. The following elaboration clarifies the figure, with added simplified examples for better understanding:

1. **Making data searchable**: This stage is initiated way before any request is received. The goal of this stage is to *properly* store data and documents so they can be retrieved at a later stage when a request is received.
2. **Collect relevant documents**: In this stage, documents stored in the previous step can be located using Search & Find. We split this section into the following three parts:
   Example: We will look for data about clean water sources for farm animals.
   (a) **Collect relevant vocabulary**: Based on the request, a person can manually determine relevant keywords and themes that relate to it.
   Example: Keywords or themes could be: 'clean' or 'access to'.
   (b) **Make vocabulary more concrete**: Based on the identified themes and terms, a person needs to refine the vocabulary by manually identifying related words.
   Example: If we look at water, it could be 'clean', 'access to', or 'supply'. If we look at farm animals, we need to take all sorts of kinds of farm animals into consideration, like: 'cows', 'chickens', 'goats', etc..
   (c) **Create search queries**: Since Search & Find works with boolean search, the search query needs to be carefully constructed.
   Example: (water NEAR3 (clean OR 'access to')) AND ('farm animal' OR cow* OR chicken* OR goat*)*
3. **Determine confidentiality of documents**: When documents have been retrieved, some documents might not contain completely reliable data. This should be filtered out.
4. **Add metadata**: To maintain findability in the future, metadata will be added to the whole dossier.
5. **Anonymize sensitive data**: Retrieved documents might contain sensitive personal data. This should not be made publicly available.
6. **Create final dossier**: Create a publication version of the request and add the constructed dossier to publish.

To address the concerns outlined, the current workflow used by Woo coordinators is notably outdated and contributes further to operational inefficiencies. This system mainly relies on manual processing and tracking of information requests, which are prone to errors and delays. Unfortunately, even with this entire process of constructing thorough search queries (steps 2a until 2c), still a lot of unwanted data is getting retrieved and possibly a lot of data is not being retrieved. Moreover, the problem is worsened by the semantic loss in chat messages and email messages. An example of this semantic loss is where a conversation about a specific topic that has initially started

in real life, is resumed by means of chat messages or email messages. This causes potentially crucial information can be misinterpreted or lost entirely during communications. The combination of these issues not only affects the immediate retrieval of information but also makes the long-term management and preservation of governmental records difficult.

Within the government, the current limitations in document management and retrieval systems include inconsistent metadata tagging, fragmented data sources, and the usage of boolean search queries. Inconsistent metadata tagging can lead to difficulties in locating documents, while fragmented data across various platforms requires a comprehensive search capability that can integrate multiple sources. Furthermore, boolean search queries, although precise, often lack the ability to capture the nuances of language, which can result in missing relevant documents.

These traditional search protocols fall short of modern solutions, which use more advanced algorithms to deliver better results and more accurate rankings. As mentioned in Section 1.3, SSC-ICT has been working on a RAG solution. The government's current reliance on boolean search queries suggests that the change to RAG could be quite large and optimistic. However, even a smaller change to a more modern retrieval system could already significantly enhance the document discoverability within the government.

2.2. Sparse Retrieval

Sparse retrieval is a fundamental technique used in information retrieval. This approach utilizes sparse vectors in high-dimensional space, typically based on the Bag of Words (BoW) [45, 66] model. In these representations, vectors consist almost entirely of zeros, due to the fact that any given document only uses a small subset of the whole vocabulary. Therefore the name 'sparse' is given to this kind of retrieval.

Traditional sparse retrieval mainly focused on optimizing and changing the weights of the BoW representation [50], resulting in methods such as Term FrequencyInverse Document Frequency (TF-IDF) or BM25. More recent work has adopted a two-stage retrieval pipeline [8, 29], where first-stage retrieval is conducted with BoW models, and the second-stage utilizes Language Model (LM)-based reranking models [38]. Here, since LM-based reranking models are more expensive than its counterpart, the goal of the first-stage retrieval is to limit the search space for the second-stage retrieval.

Sparse retrieval methods have some advantages like explainability and relatively quick computability. Although BoW models are strong baselines [63], they also have their drawbacks. For example, they are insensitive to word order and grammatical structure [45, 52]. They suffer from the lexical gap and do not generalize well [2], as they look for literal similarity between queries and documents. To address these problems, there have been attempts to substitute the standard BoW approaches with neural rerankers such as dense retrieval rerankers [8]. By incorporating contextual semantic information, these approaches address some limitations of the BoW models.

2.2.1. Probabilistic Relevance Framework

The Probabilistic Relevance Framework (PRF) is a framework for document retrieval that utilizes probabilistic models to calculate the likelihood that a document is relevant in response to a user query. This approach incorporates a probabilistic approach into the retrieval process, providing a different formal basis for retrieval models and results in different techniques for setting term weights [32].

One of the most outstanding developments stemming from PRF is the Best Matching 25 (BM25) score function, as can be seen in Equation 2.1, which has been widely acknowledged as one of the most effective text-retrieval algorithms to date [48]. The BM25 algorithm improves on the PRF by adjusting for practical factors like document length and term frequency, ensuring that longer documents are not unfairly favored and that term frequency has a balanced impact on scoring.

The BM25 equation incorporates some important components that contribute to its effective-

ness. The term frequency $f(q_i, D)$ represents the frequency of a query term q_i within a document D. This is adjusted by the term k_1 to manage the impact of term frequency saturation. Additionally, the inverse document frequency $\text{IDF}(q_i)$ component ensures that terms common across many documents are given less weight, thereby emphasizing more discriminative terms.

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} \tag{2.1}$$

$$\begin{aligned}
\text{IDF}(q_i) &= \log\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right) \\
f(q_i, D) &= \text{frequency of } q_i \text{ in document } D \\
|D| &= \text{length of document } D \\
\text{avgdl} &= \text{average document length in the corpus} \\
k_1 &\approx 1.2 \quad \text{(tuning parameter)} \\
b &\approx 0.75 \quad \text{(tuning parameter)} \\
N &= \text{total number of documents in the corpus} \\
n(q_i) &= \text{number of documents containing } q_i
\end{aligned}$$

Nowadays, there are a lot of variants and refinements of the BM25 algorithm, each proposing adjustments to the original algorithm to cater to specific use cases. These adaptations often entail adjusting parameters related to the term frequency and document length to enhance their performance on the datasets [30]. For instance, BM25+ introduces a lower-bound term frequency component. This way, no matter how long the document is, a single occurrence of a search term contributes at least a constant amount to the retrieval status value [54]. Another variant, BM25L, modifies the length normalization factor to mitigate the over-penalization of longer documents [54].

However, despite these adaptations, there seems to be no clear evidence that one of the ranking functions is systematically better than the others [54]. Research has shown that while some adaptations may yield improvement under specific conditions, no one modification to BM25 consistently surpasses the rest of the algorithms [54].

2.3. Dense Retrieval

Dense retrieval methods are different from sparse retrieval methods by utilizing dense vectors to encode both the documents and queries in a continuous vector space [23]. Using deep learning models, they create word embeddings, phrase embeddings, and document embeddings, that map words, phrases, or documents to this vector space [13]. The main benefit of using dense retrieval over the sparse counterpart is to solve the semantic matching problem [14], as it uses neural models to obtain contextual embeddings of the corresponding documents that allow learning beyond lexical similarities [18]. These models transform a piece of text into a set of numbers in an often low-dimensional [64] vector space of predetermined size. The resulting vectors are semantically close to the text, which effectively captures the essence of the text, addressing the limitation of BoW methods. Similarity searches [20], such as cosine-similarity are then calculated between the vector space and the vectorized query, to find pieces of text that semantically are the closest to each other.

Another advantage of using dense retrieval over sparse retrieval is its speed when dealing with large-scale databases. Traditional retrieval methods face scalability issues that result in increased latency during inference, due to the limitations with high-dimensional vectors [15]. Vector databases use a method known as the Approximate Nearest Neighbor (ANN) search, which is considered quicker than the traditional k-nearest neighbor (kNN) search in terms of searching. As the name implies, ANN algorithms approximate the nearest neighbors and thus can be less

precise than a kNN algorithm. However, in many high-dimensional search applications, users can be satisfied with approximate, or incomplete results, that come close to the real result [7].

Dense retrieval based on BERT models [5], has become a common approach for information retrieval [8]. One state-of-the-art model that utilizes a modification of this BERT model is Sentence-BERT [47]. This modification focuses specifically on efficiently deriving semantically meaningful sentence embeddings. Therefore the tasks that Sentence-BERT specializes in, are large-scale semantic similarity comparison, clustering, and information retrieval using cosine-similarity [47].

One important indexing optimization for vector databases is the chunking strategy. The quality of this chunking strategy can determine whether the correct context can be obtained in the retrieval phase. The most common method to split the documents is by a fixed number of tokens [53]. Where larger chunks can capture more content, but may also generate more noise and require more processing power. Conversely, smaller chunks may not fully understand the necessary context, but they do have less noise, and are less heavy to compute [10]. Chunking leads to the truncation of sentences, which will make the retrieval process harder. Methods like recursive chunking and overlapping chunks can help combine information spread across multiple chunks, despite truncations within and between sentences, enabling a more comprehensive retrieval process.

Dense retrieval models using nearest neighbors search have shown good results [16, 27, 33, 62]. Neural networks have shown strong performance in areas and use where there are extensive amounts of training data available [18]. Creating these datasets is naturally hard, due to the extensive size of the required data to be considered sufficient. This results in that there are low amounts of models where specific domain data is available. A potential solution to this problem could be to train a general dense retriever on a large dataset, and then apply it to specific domains without additional learning [18], this is also referred to as a zero-shot setting. But unfortunately, these kinds of domain specific dense retrievers are outperformed by the classical sparse retrieval methods, which do not require supervision [52].

Another challenge in dense retrieval remains to properly avoid negative results during the representation learning [23]. The models have to distinguish relevant documents from all irrelevant ones in the entire corpus, which is critical to maintaining high retrieval accuracy [62]. Techniques such as contrastive learning can be used to mitigate and refine the embeddings, by making sure that the retrieved documents are semantically closer to the query than the documents that are irrelevant [64]. By incorporating contextual cues and semantic knowledge, models can also better understand nuanced query intents and document semantics, which helps in reducing the retrieval of false positives [16]. Incorporating more classical BoW models can still enhance its performance, especially because the dense retrieval models lack of explicit term matching [8]. By merging the different retrieval methods, the models can get the best of both worlds, by getting the desirable properties of BoW models like term matching, and inverted indices [8].

2.4. Large Language Models

Large language models (LLM) represent a significant advancement in the field of AI [24], and have influenced information retrieval by leveraging deep learning to understand and generate text. These models are trained on big amounts of data, enabling them to capture complex linguistic and semantic patterns, and are able to complete language-related tasks with high accuracy.

One key development in this area, which has shaped the way we know LLMs nowadays, is the transformer architecture described [5]. Which moved away from traditional recurrent neural networks (RNNs) and convolutional neural networks (CNNs) to a structure that relies on self-attention mechanisms [56]. This self-attention mechanism enabled the model to weigh the importance of different words in a sentence relative to each other. Since a word can have different meanings in different contexts, this self-attention mechanism allows LLMs to better understand the context and semantics.

While less capable than humans in many real-world scenarios, state-of-the-art LLMs like GPT-

4 [39] and Gemini [51] have demonstrated remarkable results on many natural language processing tasks. These models excel in tasks such as text generation, translation and even problem-solving.

Despite their strengths, LLMs also present challenges like high computational costs, slower response times, and the necessity for vast amounts of training data. Additionally, because these models often function as 'black boxes', it can be difficult to understand or explain why specific documents are retrieved, which is particularly problematic in situations where transparency and clear reasoning are crucial. This black-box nature of LLMs is especially concerning in legal contexts [13]. Without a clear understanding of how LLMs relate different texts to one another, it remains challenging to validate their performance, which can hinder their use in legal settings.

Another well-known problem is the tendency of LLMs to hallucinate. This is a phenomenon where the language model generates factually incorrect information, which is not based on their training data. In legal context, such errors can be catastrophic. However, faulty LLMs have even resulted in the misrepresentation of the law, including hallucinations of fake citations [13]. The risk of these hallucinations questions the reliability of LLMs for legal research, where even minor inaccuracies can have serious consequences. Essentially, LLMs have not yet been demonstrated to be as effective as currently recognized for legal research [13, 31].

2.5. General Information Retrieval

2.5.1. Dutch/Multilingual Information Retrieval

The application of dense retrieval models across languages has significant problems due to linguistic diversity and varying semantics. The studies on these models have almost solely been performed on the English language [19]. The models, which rely on understanding and processing natural languages, must be adapted to accurately capture and interpret the different syntactic and semantic characteristics in different languages. Therefore it is necessary to develop specialized techniques or adaptations to the current existing model, to ensure similar performance in different languages.

Multilingual BERT (M-BERT) is an extension of the original BERT model [5], which was initially designed to handle tasks in only the English language [57]. However, M-BERT is pre-trained on the text from Wikipedia in 104 different languages and fine-tuned using task-specific supervised training data from one language and evaluated in a different language. Results show that M-BERT is able to perform cross-lingual generalization well, as the experiments show that high lexical overlap between languages performs well. M-BERT is also able to perform well between languages with zero lexical overlap. While M-BERT does create multilingual representations, certain language pairs are processed better than others due to the limitations in the different representations [44].

A significant advancement in Dutch information retrieval has been the introduction of BERTje [57], which is a BERT-based model specifically pre-trained on a Dutch language corpus. Although M-BERT is also trained on the Dutch language, BERTje has shown to consistently outperform the equally-sized M-BERT model on a variety of Dutch language tasks, including entity recognition, sentiment analysis, and, crucially, information retrieval [57].

2.5.2. Preprocessing

Common preprocessing techniques to improve the results in sparse retrieval include lemmatization, tokenization, and stopword removal. While these techniques may lose some detail about the text, they can significantly enhance the effectiveness of information retrieval systems by reducing the complexity and size of the text data [32]. Preprocessing data in information retrieval has naturally primarily been executed on sparse frameworks [25, 55], whereas there has been very little attention on preprocessing on dense frameworks [3].

A study by Camacho-Collados and Pilehvar [3] shows that preprocessing on neural-based frameworks can be quite important and should not be overlooked. They have run multiple experiments testing lower-casing, lemmatizing (reducing words to their base form) and multiword grouping

(grouping phrases that function as a single unit) on a CNN model, grouping on several datasets. Their results show that simple tokenization can work equally or better than more complex techniques such as the aforementioned techniques, except for domain-specific datasets, in which sole tokenization performs poorly. In a study by Rahimi and Homayounpour [46], they mention that removing stop words improves the word vectors for finding analogy relations and similarities between verbs. Therefore, it is important to keep the stop words for better results. However, there is no one-size-fits-all solution to this issue; the optimal approach will always depend on the specifics of the problem [3, 46].

2.5.3. Information Retrieval in Noisy Data

We define noisy data as data that is incomplete or corrupted. These can often be obtained through technologies such as automatic speech recognition (ASR) or optical character recognition (OCR). Such imperfections significantly complicate the retrieval process by impacting the accuracy and reliability of the information retrieved [12, 34]. One of the reasons it can complicate the retrieval process is the significant increase in the number of unique index terms, which may include many incorrect terms in noisy data [34]. The presence of noise demands robust preprocessing and normalization techniques, alongside sophisticated methods such as query expansion and enhanced matching measures, to maintain effective retrieval performance. The normalization and preprocessing of noisy data can significantly reduce the variability that hinders effective retrieval. Techniques like stemming are essential for minimizing the impact of noisy data by reducing the data complexity and focusing on the core content that enhances retrieval performance [12]. A quantitative impact of noise on retrieval systems was demonstrated in a study by Savoy and Naji [49], where a mere 5% error rate in data, caused by OCR, resulted in an approximately 17% decrease in retrieval effectiveness, even when using diverse models such as sparse and dense retrieval methods. This degradation becomes more pronounced with higher noise levels; at a 20% error rate, the performance drops by as much as 46%. These findings highlight the critical need for adjusting indexing techniques and ranking algorithms to accommodate the variability introduced by noisy data, ensuring that information retrieval systems remain robust and effective under such conditions [12, 49].

2.6. Evaluation

Evaluation is an essential aspect of information retrieval systems, as it helps determine the effectiveness and efficiency of the system in meeting user needs. We will use various metrics and methodologies to assess how well the system retrieves relevant information in response to the request queries.

2.6.1. Precision and Recall

Precision measures the proportion of retrieved documents that are relevant to the user's query. It is defined as:

$$\text{Precision} = \frac{\text{Number of Relevant Documents Retrieved}}{\text{Total Number of Documents Retrieved}} \tag{2.2}$$

High precision means that most of the documents retrieved by the system are relevant, which is particularly important in scenarios where users expect highly accurate results, such as legal searches.

Recall measures the proportion of relevant documents retrieved out of the total number of relevant documents available in the corpus. It is defined as:

$$\text{Recall} = \frac{\text{Number of Relevant Documents Retrieved}}{\text{Total Number of Relevant Documents in the Corpus}} \tag{2.3}$$

High recall is crucial in situations where it is important to retrieve as many relevant documents as possible, in our case with Woo requests, we try to find as many related documents as possible.

Precision-Recall Trade-off: Often, there is a trade-off between precision and recall. Retrieving more documents (i.e. increasing recall) might lower precision (i.e. decrease precision) if the additional documents are less relevant. Conversely, focusing on highly relevant documents might reduce recall by missing out on some relevant documents.

2.6.2. F1 Score

The **F1 score** is the harmonic mean of precision and recall, providing a single metric that balances the two. It is particularly useful for accounting for both false positives and false negatives. The F1 score is defined as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

A high F1 score indicates that the system is performing well in both precision and recall, making it a comprehensive metric for evaluation.

2.6.3. Mean Average Precision (MAP)

Mean Average Precision is a metric that combines precision and recall across multiple queries. For each query, the average precision (AP) is calculated as the mean of precision scores obtained after each relevant document is retrieved.

$$\text{AP} = \frac{1}{N} \sum_{k=1}^{N} P(k) \times \text{rel}(k) \tag{2.5}$$

Where:

- N: The total number of retrieved documents.
- $P(k)$: The precision at rank k.
- $\text{rel}(k)$: 1 if the document at rank k is relevant, and 0 otherwise.

MAP is then calculated by averaging these AP scores across all queries:

$$\text{MAP} = \frac{1}{|Q|} \sum_{q \in Q} \text{AP}(q) \tag{2.6}$$

Where $|Q|$ is the number of queries. MAP is particularly valuable for evaluating systems on large test sets with multiple queries, as it reflects both precision and recall over a range of thresholds.

2.6.4. Normalized Discounted Cumulative Gain (NDCG)

Normalized Discounted Cumulative Gain is another popular metric for evaluating the ranking quality of retrieved documents. It takes into account the relevance of documents as well as their positions in the ranked list. The core idea is that documents appearing higher in the ranking should have more influence on the score, especially if they are highly relevant. NDCG is computed as follows:

- First, calculate the **Cumulative Gain (CG)** at each rank position:

$$\text{CG}_p = \sum_{i=1}^{p} \text{relevance}(i) \tag{2.7}$$

- Then, discount the gain based on the rank position to get **Discounted Cumulative Gain (DCG)**:

$$\mathrm{DCG}_p = \sum_{i=1}^{p} \frac{\mathrm{relevance}(i)}{\log_2(i+1)} \tag{2.8}$$

- Finally, normalize the DCG by the ideal DCG (IDCG), which is the DCG score for the ideal ranking of documents, to obtain NDCG:

$$\mathrm{NDCG}_p = \frac{\mathrm{DCG}_p}{\mathrm{IDCG}_p} \tag{2.9}$$

NDCG values range from 0 to 1, with 1 indicating a perfect ranking. This metric is particularly useful in scenarios where the order of retrieved documents significantly impacts the user experience, such as in search engines or recommendation systems.

3

Related Works

In the context of the Woo and the government as a whole, effective information retrieval is obviously important. The challenge is to ensure that government officials can access relevant and reliable information efficiently, perhaps through the tooling of other frameworks. This section presents some of the alternatives and additions that can deal with these challenges, but are not taken into account in this thesis.

3.1. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is a framework designed to add on the areas where LLMs have limitations. Currently, the ability of LLMs to access and precisely manipulate data is still limited [26]. RAG models combine parametric and non-parametric memory for language generation, which tackles many of the LLMs limitations. More explicitly, RAG integrates the strengths of traditional information retrieval methods with the generative capabilities of LLMs. The framework operates by retrieving relevant information from an external (vector) database or knowledge source, which is then used to inform and guide the generation process of the LLM. This approach allows the model to access a broader and more up-to-date knowledge base without the need for constant retraining, as it can pull in real-time data from the external database. This is particularly advantageous in domains where knowledge is rapidly evolving, or where specific, up-to-date information is required for accurate text generation.

This approach can be particularly beneficial in the context of Woo, where current information retrieval processes often result in large volumes of documents that need to be manually reviewed. In the current setup, Woo coordinators are responsible for sifting through extensive sets of documents to determine which ones are relevant to a specific request. By utilizing RAG, this manual process can be streamlined, as the LLM can retrieve, process, and summarize relevant information from numerous documents, reducing the need for human intervention at the early stages. This can save significant time and resources, making it a more efficient way to handle large-scale information requests.

However, deploying RAG in a governmental setting like for Woo requires careful consideration. While it can automate much of the retrieval and filtering process, it is essential that Woo coordinators remain precise and alert. LLMs, despite their strengths, may not always present information with complete accuracy or could omit critical details. Human oversight is still necessary to ensure that key information is not lost or misinterpreted.

3.2. Hypothetical Retrieval

A different way approach in dense retrieval systems is the use of hypothetical retrieval. HyDE [9] introduces an innovative approach in information retrieval by leveraging hypothetical answers generated by language models. Given a query, HyDE first executes a zero-shot prompting strategy to instruct a language model to generate a hypothetical answer. Since this answer is not based on any data, it is safe to assume that there is a high likelihood that this answer is hallucinated. However, this answer is intended to capture relevant patterns and semantic structures similar to the real answer, even if it contains inaccurate details. The core idea is that the hypothetical answer encapsulates the context and nuances of the query, that would be missed when comparing the query with the vector database directly.

Once the hypothetical answer is generated, it is encoded into an embedding vector using a suitable embedding model. Instead of the original query embeddings, this embedding will be used to perform similarity searches within a vector database. By comparing the embedding of the hypothetical answer to those in the database, the system identifies a neighborhood in the embedding space where relevant documents are likely to reside. Evaluations have demonstrated that HyDE outperforms traditional retrieval methods in specific contexts.

3.3. Contextual Retrieval

Another weakness of the retrieval in dense retrieval systems is the loss of context between chunks. Typically, documents are divided into chunks with chunk overlap to help the processing and reduce noise, but this can lead to the lack of sufficient contextual information that might be relevant for a chunk. For example, the main topic of a page might be introduced at the beginning, while subsequent sentences continue discussing the topic without explicitly mentioning it. When these later chunks are processed independently, they may not appear relevant to the query, despite containing related information.

The contextual retrieval method proposed by Anthropic [17], ensures that each chunk of information retains its surrounding context, by prepending the context to every chunk, ensuring that all the chunks are context aware. This approach transforms each chunk into a more comprehensive representation of the document, enhancing the retrieval system's ability to understand the context.

The results presented in the paper on this method have been promising. Contextual embeddings generated using this approach have been shown to reduce the top-20-chunk retrieval failure rate by 35% [17].

4

Data Preparation

To perform improvements on steps 1 through 8 depicted in Figure 1.1 and to address our research questions, we need to conduct experiments using a well-defined dataset of government documents and associated queries. For meaningful analysis, it is crucial to have a combination of documents that can serve as ground truth in our experiments. This way we can use the Woo request files as input queries and the corresponding answer files as the ground truth.

For this experiment, we have selected data from Woogle[1], using a data dump taken on April 1, 2024. A more recent version of this data dump (April 19, 2024) can be found on DANS Data Station Social Sciences and Humanities[2]. Woogle is an external party (i.e. not affiliated with the Dutch Government) run by researchers at the University of Amsterdam that aims to make all government info reusable. This means that they collect all published documents about the Woo from every party, including the aforementioned request files and answer files. To narrow down the scope of the project, we will limit our experiments to using data provided by a select few Dutch ministries only.

There are several reasons for the selection of data from Woogle for this experiment, rather than accessing real-time government data directly from their databases. Firstly, the nature of the data available through government channels often includes sensitive or confidential information, such as personal data, which would raise concerns about privacy and ethical considerations. The data provided by Woogle only contains data that is already available to the public. Data that can be considered sensitive or confidential would already be redacted in some form by this point.

Secondly, the structure and organization of government data yield different challenges as well. Data is typically spread across multiple databases, each maintained by different ministries, and might have lots of different forms. This makes it very difficult to collect and process all the data, but also to process the data in such a way that it is uniform over the whole experiment. In contrast, the data from Woogle has one coherent structure, split up over multiple databases. The uniform structure makes it easy to work with and combine the databases (see: section 4.1), and will simplify the process of data retrieval and analysis.

Lastly, the Woogle database also shows the correlation between files, which we will use as ground truth. As mentioned before, the data in Woogle only contains data that is already available to the public. This means that the data is in the form of a *publication version*, which entails a full **dossier**. A dossier has a **request** file, in which a publication version is shown of the original request that is received by the ministry. Aside from this request file, one or more attachments should be included to the dossier. These extra attachments are the documents or files that the

[1] `https://woogle.wooverheid.nl/`

[2] `https://ssh.datastations.nl/dataset.xhtml?persistentId=doi:10.17026/dans-zau-e3rk`

Table 4.1: Data Types in Woogle Databases

(a) Dossier Database

Dossier Data Types	
Field	**Data Type**
dc_identifier	String
dc_title	String
dc_description	String
dc_type	String
foi_type_description	String
dc_publisher_name	String
dc_publisher	String
dc_source	String
foi_valuation	String
foi_requestText	String
foi_decisionText	String
foi_isAdjourned	String
foi_requester	String

(b) Document Database

Document Data Types	
Field	**Data Type**
dc_identifier	String
foi_dossierId	String
dc_title	String
foi_fileName	String
dc_format	String
dc_source	String
dc_type	String
foi_nrPages	Integer

(c) Bodytext Database

Bodytext Data Types	
Field	**Data Type**
dc_publisher_name	String
dc_publisher	String
foi_documentId	String
foi_pageNumber	Integer
foi_bodyText	String
foi_bodyTextOCR	String
foi_hasOCR	Boolean
foi_redacted	Float
foi_contourArea	Float
foi_textArea	Float
foi_charArea	Float
foi_percentageTextAreaRedacted	Float
foi_percentageCharAreaRedacted	Float
foi_nrWords	Integer
foi_nrChars	Integer
foi_nrWordsOCR	Integer
foi_nrCharsOCR	Integer

requester has asked for. Therefore, a request file could be used as input, with the corresponding attachments as ground truth.

4.1. Overview of Data Fields

Table 4.1 provides a comprehensive overview of all the data fields that are present in our dataset. The data model follows a hierarchical approach, starting from the highest level: dossiers, to documents, and finally bodytext, which represents the individual pages.

The integration and consistency across these different levels of data are achieved through the use of unique identifiers, which serve as relational keys that connect the various tables. Specifically, the 'dc_identifier' field in the dossier data type corresponds directly to the 'foi_dossierId' in the document data type. Similarly, the 'dc_identifier' field used in the document data type aligns with the 'foi_documentId' in the data type. This use of identifiers across different levels of data makes it possible to merge the tables into one big database. For convenience, we will merge the databases into one single database for the experiments. These steps help us get the data ready so it is organized and clean for our analysis.

While merging, we only keep the fields that are necessary for the experiments. For convenience, some fields are added to ensure a unique ID. Here is a breakdown of each field that we keep and its relevance:

1. page_id
   This is a concatenation of the corresponding 'foi_documentId' with 'foi_pageNumber'
2. document_id
3. dossier_id
4. bodyText
   This includes the primary content of each document. We use the 'foi_bodyText' data where possible. Otherwise, 'foi_bodyTextOCR' is taken.
5. type
   Indicates the type of document (i.e. request, attachment, or decision), which can be used to distinguish the documents in the dataset.
6. publisher
   Includes from which ministry the data is from. Can be used to specifically filter data per ministry.
7. source
   Includes the URL of the original document. Can be used to find the original data, in case we manually want to check something.

4.2. Data Quality

The quality of data retrieved from Woogle still presents several challenges that need to be addressed, specifically due to noise found in both the bodytext and bodytextOCR fields.

The primary reason why the data is noisy and not in a clean, ready-to-use format, is that all the documents have undergone a series of transformations before it has been published. Since confidentiality is a big concern in these documents, personal data must be redacted before it can be published, several measurements have been taken. Right now, documents are manually redacted on the computer, but to ensure that it is safe, the pages get printed and then scanned. This scanned product is subsequently processed using OCR to extract the textual information.

Unfortunately, this process can introduce numerous errors and inconsistencies. For example, OCR might introduce incorrect characters or missing text or it might convert images (such as logos)

to text among many other things. OCR errors are also common when dealing with low-resolution scans or when pages have complex formatting. Such as tables or multi-column layouts.

The difference between bodytext and bodytextOCR lies in the stage and tool used for OCR. Bodytext is generated using OCR performed by the government, which produces directly readable text from the files. In contrast, bodytextOCR is processed by Woogle, leading to potential discrepancies between the two versions.

The data should be appropriately cleaned and pre-processed to minimize the noise in the data. However, no matter the preprocessing, without a large amount of manual intervention, the data will never match the quality of the original documents.

5

Methodology

In this section, we will provide the outlines of the experiment. We will query the dataset described in Section 4 using different models and assess their performance. Additionally, we will also apply different preprocessing techniques to both the query files and the ground truth files and compare the evaluation of those. With this, we can see if these different techniques can help with steps 1 through 8 in the RAG design used by SSC-ICT as depicted in Figure 1.1.

5.1. Experiment

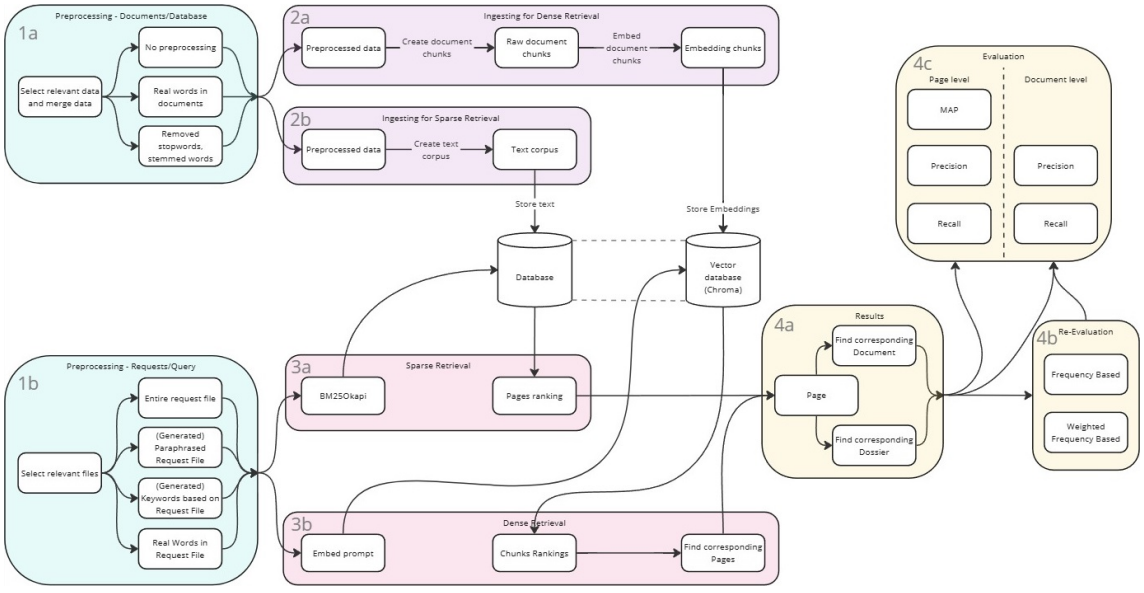

Figure 5.1: Flow design for the experiments.

We will conduct a quantitative experiment using the dataset provided by Woogle, prepared as outlined in Section 4. All experiments are running locally on an MSI Titan 18 HX A14VIG using a GeForce RTX 4090 GPU with 16 GB VRAM and a 24-core Intel Core i9-14900HX. We will experiment with a limited set of Dutch Ministries. We will conduct experiments using a select group of Dutch ministries, specifically chosen to evaluate potential differences in performance across varying datasets. The ministries included in this experiment are:

1. Ministry of the Interior and Kingdom Relations (in Dutch: Ministerie van Binnenlandse Zaken en Koninkrijksrelaties)

   Chosen as a reference point as SSC-ICT is part of this ministry.

2. Ministry of General Affairs (in Dutch: Ministerie van Algemene Zaken)

   Selected for its smaller dataset.

3. Ministry of Foreign Affairs (in Dutch: Ministerie van Buitenlandse Zaken)

   Included to allow comparison with the Ministry of the Interior, as they seem to be quite similar.

4. Ministry of Finance (in Dutch: Ministerie van Financiën)

   Selected to observe its performance on data that includes more numerical information.

5. Ministry of Justice and Security (in Dutch: Ministerie van Justitie en Veiligheid)

   Selected for its relatively larger dataset.

The goal of these experiments is to explore and evaluate different text retrieval techniques to determine the most effective approach for matching requests with corresponding documents within a dataset.

An outline of the experiment can be found in Figure 5.1. To describe the graph in a structured and clear way, we have split up the description into sections. A summary of the steps is provided below:

- **Sub-figure 1:** preprocessing, Subsection 5.1.1
  - **Sub-figure 1a:** describes the preprocessing of the documents/ground truth in 3 different ways.
  - **Sub-figure 1b:** describes the preprocessing of the query files in 4 different ways.
- **Sub-figure 2:** ingesting (database creation), Subsection 5.1.2
  - **Sub-figure 2a:** describes the database creation for dense retrieval/embeddings.
  - **Sub-figure 2b:** describes the database creation for sparse retrieval (in this case BM25).
- **Sub-figure 3:** retrieval, Subsection 5.1.3
  - **Sub-figure 3a:** describes the retrieval for the sparse approach.
  - **Sub-figure 3b:** describes the retrieval for the dense approach.
- **Sub-figure 4:** results and evaluation, Subsection 5.1.4
  - **Sub-figure 4a:** describes the results and the relation to the documents and dossiers.
  - **Sub-figure 4b:** describes the the frequency-based re-evaluation as described in Subsection 5.1.4.1.
  - **Sub-figure 4c:** describes all the final evaluation metrics.

5.1.1. Preprocessing

The initial steps involve the preprocessing of data, depicted by the blue squares.

Figure 5.1, sub-figure 1a represents preprocessing of the data that we will store in the database. We process this data in 3 different ways, which will result in 3 distinct databases:

1. No preprocessing
2. Keeping real words

   Given that the data is generated through OCR and contains considerable noise, including misspelled words, we include a 'real words' preprocessing option. This method filters out any text that is not recognized as a valid Dutch word based on the following list[1], which is a list that includes every word in every form.
   This preprocessing step is based on the assumption that sparse retrieval methods are unlikely to benefit from incorrectly recognized OCR words. Additionally, by removing these non-words, we aim to improve the semantic quality for dense retrieval, as irrelevant or hard-to-interpret words are removed, allowing the model to better capture the intended meaning.
3. Stop word removal and stem words

   This preprocessing method, is frequently used for sparse retrieval, for example in [22, 61]. Stop word removal eliminates common words such as 'and', and 'the' which are often considered non-informative for retrieval tasks. Stemming reduces words to their basic counterpart (e.g. 'runner' and 'running' will get reduced to 'run'), to standardize variations of the same word. While this type of preprocessing is less common in dense retrieval due to the potential loss of semantic meaning, we have applied it across both sparse and dense retrieval in this study. This allows for a consistent preprocessing baseline, allowing us to compare effects across different retrieval methods.

Figure 5.1, sub-figure 1b represents preprocessing of the input data (i.e. the query). We process this data in 4 different ways, which will result in 4 types of query files:

1. No preprocessing
2. Generated paraphrased request file

   A request file can contain multiple pages, and the essence of the request is often a paragraph on the first page. This means that there is a lot of unwanted noise in the rest of the document. Here, we run a LLM to extract the main intent of the whole document. For this experiment, we use Llama 3[2] to extract a concise summary. By reducing each request file to a few key sentences, we aim to streamline the data for dense embedding algorithms. This representation minimizes the noise, allowing the embedding model to capture the essence of the request more effectively, which should increase the performance.
3. Generated keywords based on request file

   To further enhance sparse retrieval, we use a LLM to extract key terms from each request file, that should capture the topics and intent of the document. By focusing on keywords, we can create a more targeted representation of the requests in which BM25 can work more effectively.
4. Real words in request file

   Just like in the database, the query files have also been generated through OCR. For the same reasons as with the database, we will filter out all words that are not proper Dutch words.

[1] https://github.com/OpenTaal/opentaal-wordlist/blob/master/elements/wordlist-ascii.txt
[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B

5.1.2. Ingest

When creating the database that is used in sparse retrieval, as depicted in Figure 5.1, sub-figure 2b , the data undergoes a straightforward process. After the preprocessing steps, the data is directly compiled into a text corpus without further transformation. This text corpus contains the original structure of the documents, which preserves term frequencies.

For creating this vector database, we need to chunk the data before embedding it, as depicted in Figure 5.1, sub-figure 2a. We use ChromaDB[3] as vector database, with a Langchain[4] abstraction layer. It is imperative to split the data into smaller, manageable, units before generating their embeddings. To split the data, we use the NLTK [28] text splitter, with chunk_size = 1024. Embedding models have limitation have limitations on the maximum input sequence [5], and exceeding this length can lead to truncation of the data which is not preferable. Furthermore, chunking enhances the semantic representation captured by the embeddings. Smaller chunks can enable the model to focus on specific contents in the data.

For the experiment, we have decided on 2 different embedding models. Namely, GroNLP/bert-base-dutch-cased[5][57] and sentence-transformers/all-MiniLM-L6-v2[6][58]. They are different because they excel in different natural language processing tasks.

Bert-base-dutch-cased is a fill-mask model, which means that it is suited for word-level corrections, but it is not suited for understanding the meaning of a full sentence. This makes it less ideal for tasks like semantic search in vector databases, where the focus is on understanding the overall context or intent of a query rather than just individual words. However, its key strength for our use case lies in the fact that it is trained on Dutch data. Since our database is entirely in Dutch, this model is likely to have a better understanding of the language, including expressions and grammatical structures that other, more general models may miss. By leveraging this Dutch-language training, it can provide more contextually accurate token-level predictions, which can be a valuable asset for certain types of analysis within the database.

All-MiniLM-L6-v2 is a sentence similarity model, which means that it is designed to understand and compare the overall meaning of sentences, rather than individual words. This makes it ideal for tasks like semantic search, where the ability to capture broader context is crucial. However, the model is only trained on English data, which makes it work limited for Dutch text. However, the model might understand similar sentence structures that are comparable between the English and Dutch languages.

5.1.3. Retrieval

Figure 5.1, sub-figure 3a shows how to retrieve with sparse retrieval. This can be executed quite straightforwardly, as the preprocessed data can be directly queried. After the queries have been executed we receive a ranking of the pages.

Figure 5.1, sub-figure 3b shows how to retrieve with dense retrieval. In this case, we first need to embed the prompt. The same embedding function as for the ingestion has to be used. This embedding will then be queried in the vector database, and the most similar chunks will be retrieved. More precisely, the chunks that are semantically most similar to the input query will be returned. Lastly, since we retrieve chunks instead of pages, we still need to find the corresponding pages based on the chunks.

To compare the performance over different amounts of data, we have decided to retrieve different amounts of data in the retrieval step, ranging from 10 to 100 pages in increments of 10.

[3] https://github.com/chroma-core/chroma
[4] https://www.langchain.com/
[5] https://huggingface.co/GroNLP/bert-base-dutch-cased
[6] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

5.1.4. Results and Evaluation

After having retrieved all the pages, we can extract the corresponding document and corresponding dossier depicted in Figure 5.1, sub-figure 4a. This is a simple database lookup where we get the document_id and dossier_id from the database.

We can evaluate our results using various metrics, such as Precision, Recall, and MAP as can be seen in Figure 5.1, sub-figure 4c. In the context of our application, we have decided to present these metrics while keeping in mind that some are more important to our objectives than others.

Precision measures the proportion of retrieved documents that are relevant, which shows the accuracy of the retrieval system. Naturally, precision is important because it indicates how accurate the retrieved files are. However, the main drawback of relying solely on precision is that it does not account for how many of the total relevant documents are retrieved; it does not measure the system's ability to find all relevant files.

This is where recall is more relevant, as recall measures the proportion of relevant documents that are retrieved out of all relevant documents available. In the context of Woo, it is vital that we retrieve as many relevant documents as possible. Missing relevant documents could lead to incomplete information for the user, which may have dire consequences. Therefore, even if increasing recall leads to a decrease in precision, by retrieving more documents overall, including some irrelevant onesthe trade-off is arguably acceptable in our case.

Ultimately, we face a trade-off between precision and recall. To evaluate this trade-off comprehensively, we can use MAP. MAP calculates the average precision for each query and then takes the mean across all queries, effectively combining precision and recall into a single metric. This provides a more holistic view of the system's performance over a range of recall levels and is useful for comparing different retrieval systems. This metric is also less sensitive in choosing how many documents to retrieve because later retrieved documents are less relevant for this metric.

Similar metrics like the F1-score, which combines precision and recall, are not being evaluated. This metric assumes that precision and recall are equally important. Since our application prioritizes recall over precision, this metric does not reflect our goals.

Instead, we focus on recall and MAP as our primary metrics while still considering precision to understand the impact on accuracy. By emphasizing recall, we ensure that our system retrieves as many relevant documents as possible, aligning with the critical needs of our users in the context of Woo.

5.1.4.1. Frequency-Based Re-evaluation

We introduce two additional methods for assigning relevance scores to the retrieved pages, which can be seen in Figure 5.1, sub-figure 4b. Namely, frequency-based re-evaluation and weighted frequency-based re-evaluation, which is an enhancement on the original one. These methods aim to further enhance the accuracy of the results that are being retrieved.

The frequency-based re-evaluation assigns relevance scores to documents, solely based on the frequency of their occurrences within the top n retrieved pages. The underlying assumption is that documents appearing more frequently are more likely to be relevant.

For a set of retrieved documents $\{D_1, D_2, \ldots, D_n\}$, the relevance score $S(D)$ for a document D is calculated as:

$$\mathrm{S}(D) = \frac{\text{Frequency of } D}{n} \tag{5.1}$$

The following demonstrates a simple example of the mentioned method. Consider a retrieval scenario where $n = 4$ documents are retrieved in the following order:

1. Document A
2. Document B
3. Document C

4. Document A

The corresponding frequencies are:

1. Document A: 2 occurrences
2. Document B: 1 occurrence
3. Document C: 1 occurrence

Applying the frequency-based re-evaluation we get:

$$S(A) = \frac{2}{4} = 0.5$$
$$S(B) = \frac{1}{4} = 0.25$$
$$S(C) = \frac{1}{4} = 0.25$$

Thus Document A has the highest relevance score.

The weighted frequency-based re-evaluation method extends the original method by incorporating the positions of the documents into the relevance scoring. This method gives documents that appear later in the retrieval list are less relevant to ones appearing earlier.

For each occurrence of a document D at rank k, we calculate a weighted contribution to the relevance score based on the precision at that rank. The overall relevance score $S(D)$ is the sum of these contributions. The relevance score is derived from the formula of average precision as shown in equation 2.5:

$$\mathrm{S}(D) = \frac{1}{N} \sum_{k=1}^{N} P(k) \times \mathrm{rel}(k) \tag{5.2}$$

Using the same scenario as before, we demonstrate another simple example with the other method:

$$S(A) = \left(\frac{1}{4} \times 1 \times 1\right) + \left(\frac{1}{4} \times 0.5 \times 1\right) = 0.5 + 0.25 = 0.375$$
$$S(B) = \frac{1}{4} \times 0.5 \times 2 = 0.25$$
$$S(C) = \frac{1}{4} \times 0.333 \times 2 \approx 0.166$$

When evaluating these methods, we can organize the retrieved pages into four brackets:

- True Positive (TP): The page is part of the ground truth, and the re-evaluation has also assigned it to be relevant.
- False Positive (FP): The page is **not** part of the ground truth, but the re-evaluation has assigned it to be relevant.
- True Negative (TN): The page is **not** part of the ground truth, and the re-evaluation has **not** assigned it to being relevant.
- False Negative (FN): The page is part of the ground truth, but the re-evaluation has also assigned it to it being **not** relevant.

Here, we try to maximize TP and TN and minimize FP and FN, to ensure the most accurate identification of relevant pages and irrelevant pages.

5.2. Time Taken

In the context of information retrieval, it is important that processes happen in a timely manner, so that the user can get the result fast. Evaluating the time taken at each stage of the experiment is important to understand the overall efficiency of our approach. If constructing a vector database is excessively time-consuming, it may hinder scalability and limit its practicality in real-world applications. Moreover, retrieval speed is even more important, as users typically expect near-instant responses, and delays can quickly lead to dissatisfaction.

We will track how long it takes to:

1. The time required for preprocessing the data and query files.
2. The time required for creating the vector stores and normal databases.
3. The time required to retrieve documents based on the request file.

The amount of data that we have available for all the ministries combined are 1253 dossiers (equals the number of input files) and 271844 A4-sized pages. The ministries individually contain:

- Ministry of the Interior and Kingdom Relations: 241 dossiers, 54207 pages
- Ministry of General Affairs: 26 dossiers, 1824 pages
- Ministry of Foreign Affairs: 196 dossiers, 30352 pages
- Ministry of Finance: 357 dossiers, 63611 pages
- Ministry of Justice and Security: 433 dossiers, 121850 pages

We will report these times as both cumulative totals across all ministries and also averaged per dossier.

6

Results

The results presented in this chapter will be the averaged results over the ministries mentioned in Chapter 5. If some part of a particular ministry presents remarkable results, this will also be highlighted. All evaluation results for every ministry individually can be found in Appendix A, on the frequency-based and weighted frequency-based method in Appendix B and on time taken in Appendix C.

6.1. Evaluation

The averaged results for MAP, precision, and recall across the different methods are shown in the tables below, with the highest value in every column highlighted. To visualize some of the performance trends, we have also included graphs for MAP. Detailed tables and graphs for MAP, precision, and recall are included in Appendix A.

Based on the results, we can determine that BM25 is the best algorithm across all metrics using this data and these evaluation metrics, which directly addresses RQ1. When comparing it to any of the dense retrieval models, we notice that BM25 performs significantly better, in some cases, outperforming it by as much as a factor of 10.

A consistent trend that we can see over all the results is that **non-processed databases** generally perform best across all methods. However, for the query files, we can see that **non-processed query files** tend to work the best for sparse retrieval methods, but **paraphrased query files** work better for the dense models.

In terms of metrics, we observe an inverse relationship between the number of retrieved documents and performance for MAP and recall. As more documents are getting retrieved, the MAP generally decreases, while the recall is getting higher and higher.

One interesting finding is that when we make the query shorter (i.e. query = keywords or query = paraphrase), the performance on BM25 gets significantly lower. However, for dense retrieval models, this is not the case. Here, the length of the query file does not seem to directly impact the results.

Table 6.1: **Evaluation Metrics** for **BM25** on **all ministries averaged**

Query File	Database	MAP@10	MAP@50	MAP@100	Precision@10	Precision@50	Precision@100	Recall@10	Recall@50	Recall@100
raw	raw	**0.7536**	**0.5396**	**0.4158**	**0.7774**	**0.5736**	**0.4515**	**0.2484**	**0.4856**	**0.5921**
	stop words	0.7469	0.5308	0.4049	0.7751	0.5670	0.4424	0.2468	0.4767	0.5791
	real words	0.6725	0.4688	0.3516	0.7055	0.5127	0.3966	0.2015	0.4079	0.5027
keywords	raw	0.2092	0.1204	0.0853	0.2465	0.1632	0.1252	0.0704	0.1353	0.1763
	stop words	0.1636	0.0970	0.0718	0.1960	0.1358	0.1078	0.0532	0.1064	0.1341
	real words	0.1527	0.0831	0.0586	0.1864	0.1223	0.0945	0.0526	0.1054	0.1346
paraphrase	raw	0.2911	0.1570	0.1067	0.3381	0.2128	0.1581	0.0993	0.1859	0.2260
	stop words	0.2394	0.1312	0.0917	0.2890	0.1850	0.1417	0.0835	0.1615	0.2002
	real words	0.1970	0.0998	0.0676	0.2408	0.1481	0.1115	0.0718	0.1319	0.1606
real words	raw	0.6658	0.4607	0.3453	0.6967	0.5029	0.3889	0.1971	0.3947	0.4860
	stop words	0.6503	0.4326	0.3187	0.6844	0.4783	0.3645	0.1851	0.3783	0.4603
	real words	0.6834	0.4792	0.3594	0.7125	0.5197	0.4013	0.1991	0.4114	0.5069

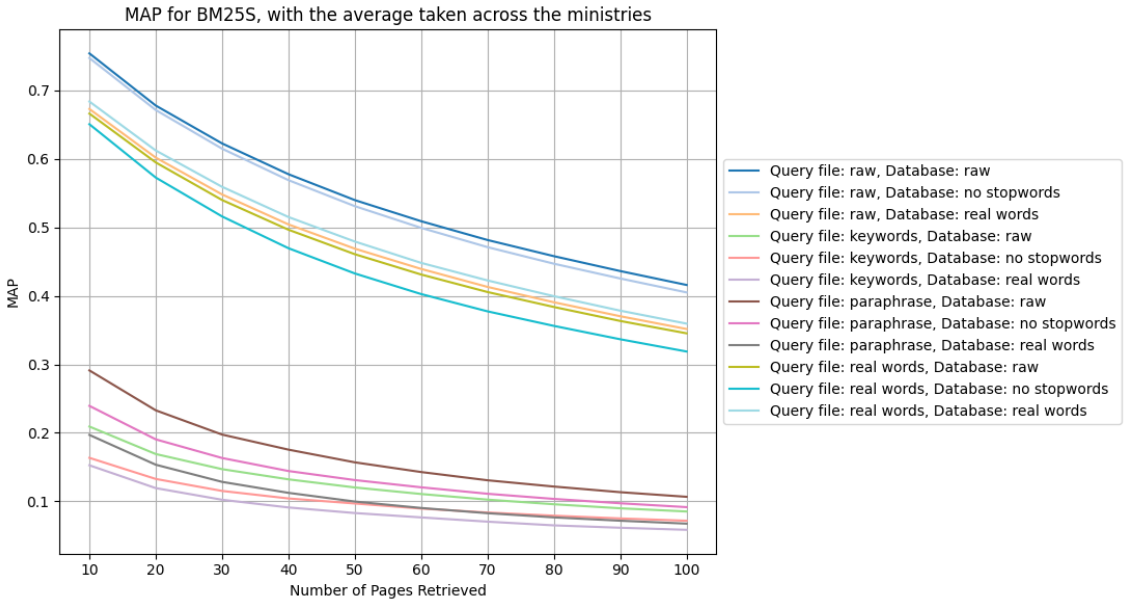

Figure 6.1: MAP for BM25, with the average taken for every ministry.

Table 6.2: **Evaluation Metrics** for **BERTje** on **all ministries averaged**

Query File	Database	MAP@10	MAP@50	MAP@100	Precision@10	Precision@50	Precision@100	Recall@10	Recall@50	Recall@100
raw	raw	**0.0585**	**0.0249**	0.0189	0.0726	0.0415	0.0288	0.0195	0.0430	0.0532
	stop words	0.0131	0.0068	0.0047	0.0207	0.0161	0.0153	0.0029	0.0099	0.0241
	real words	0.0203	0.0091	0.0064	0.0329	0.0260	0.0216	0.0157	0.0323	0.0419
keywords	raw	0.0217	0.0103	0.0094	0.0362	0.0282	0.0235	0.0162	0.0274	0.0360
	stop words	0.0080	0.0036	0.0026	0.0146	0.0135	0.0134	0.0016	0.0082	0.0170
	real words	0.0172	0.0097	0.0074	0.0351	0.0266	0.0235	0.0150	0.0246	0.0334
paraphrase	raw	0.0536	0.0231	**0.0200**	**0.0813**	**0.0499**	**0.0368**	**0.0245**	**0.0439**	**0.0533**
	stop words	0.0106	0.0043	0.0027	0.0174	0.0141	0.0127	0.0017	0.0078	0.0165
	real words	0.0280	0.0111	0.0080	0.0474	0.0314	0.0267	0.0189	0.0329	0.0463
real words	raw	0.0268	0.0127	0.0114	0.0419	0.0269	0.0215	0.0135	0.0284	0.0409
	stop words	0.0143	0.0064	0.0041	0.0216	0.0161	0.0145	0.0034	0.0084	0.0160
	real words	0.0442	0.0180	0.0114	0.0570	0.0350	0.0275	0.0191	0.0382	0.0489

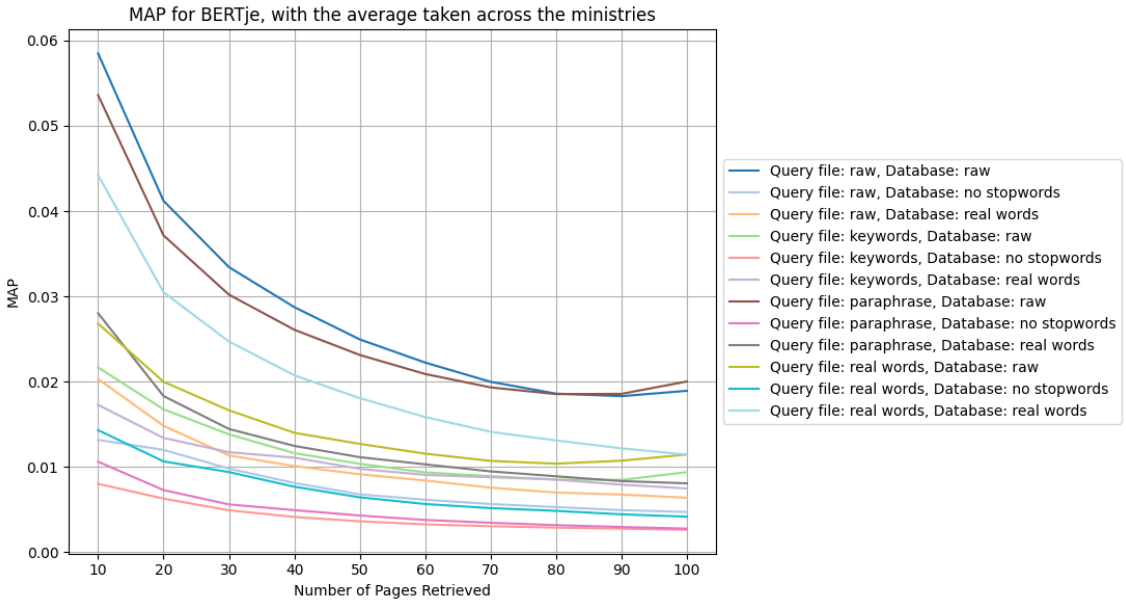

Figure 6.2: MAP for BERTje, with the average taken for every ministry.

Table 6.3: **Evaluation Metrics** for **MiniLM** on **all ministries averaged**

Query File	Database	MAP@10	MAP@50	MAP@100	Precision@10	Precision@50	Precision@100	Recall@10	Recall@50	Recall@100
raw	raw	**0.0585**	**0.0249**	0.0189	0.0726	0.0415	0.0288	0.0195	0.0430	0.0532
	stop words	0.0131	0.0068	0.0047	0.0207	0.0161	0.0153	0.0029	0.0099	0.0241
	real words	0.0203	0.0091	0.0064	0.0329	0.0260	0.0216	0.0157	0.0323	0.0419
keywords	raw	0.0217	0.0103	0.0094	0.0362	0.0282	0.0235	0.0162	0.0274	0.0360
	stop words	0.0080	0.0036	0.0026	0.0146	0.0135	0.0134	0.0016	0.0082	0.0170
	real words	0.0172	0.0097	0.0074	0.0351	0.0266	0.0235	0.0150	0.0246	0.0334
paraphrase	raw	0.0536	0.0231	**0.0200**	**0.0813**	**0.0499**	**0.0368**	**0.0245**	**0.0439**	**0.0533**
	stop words	0.0106	0.0043	0.0027	0.0174	0.0141	0.0127	0.0017	0.0078	0.0165
	real words	0.0280	0.0111	0.0080	0.0474	0.0314	0.0267	0.0189	0.0329	0.0463
real words	raw	0.0268	0.0127	0.0114	0.0419	0.0269	0.0215	0.0135	0.0284	0.0409
	stop words	0.0143	0.0064	0.0041	0.0216	0.0161	0.0145	0.0034	0.0084	0.0160
	real words	0.0442	0.0180	0.0114	0.0570	0.0350	0.0275	0.0191	0.0382	0.0489

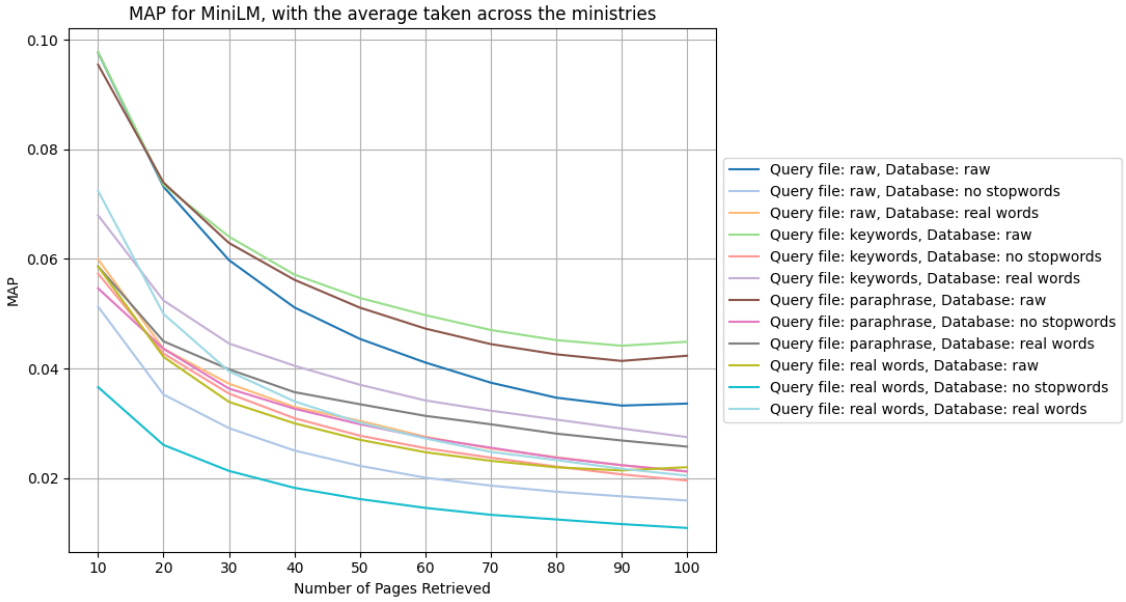

Figure 6.3: MAP for MiniLM, with the average taken for every ministry.

6.2. Frequency Based Re-evaluation

This section addresses RQ2 by coming up with a potential technological improvement by using frequency-based re-evaluation. The figures in this section present the results of the frequency-based re-evaluation. We have chosen to only show the best-performing evaluation result in this section; this means we included the best-performing result based on normal evaluation metrics, before re-evaluation. The rest of the results, with all the different retrieval methods and different numbers of pages retrieved, can be seen in Appendix B.

Figure 6.4 displays the outcomes using regular frequency for 10 number of pages retrieved using BM25 and Figure 6.5 the corresponding ROC curve and AUC. Likewise, Figure 6.6 shows the graph for the weighted algorithm with Figure 6.7 its corresponding ROC curve and AUC.

Across all the graphs, a steady decline in true positives can be observed as the threshold increases. At the same time, false positives decrease rapidly for thresholds below 0.1. This effect is particularly apparent in the weighted frequency results due to the naturally lower relevance scores compared to the standard frequency. To make these intervals clearer, we have adjusted the threshold intervals for the weighted frequencies, using 1,000 uniformly distributed intervals between 0 and 0.1, and 9 intervals between 0.1 and 1. This change allows for a more detailed analysis of the impact of smaller thresholds, especially given the more sensitive behavior of weighted frequencies at lower ranges.

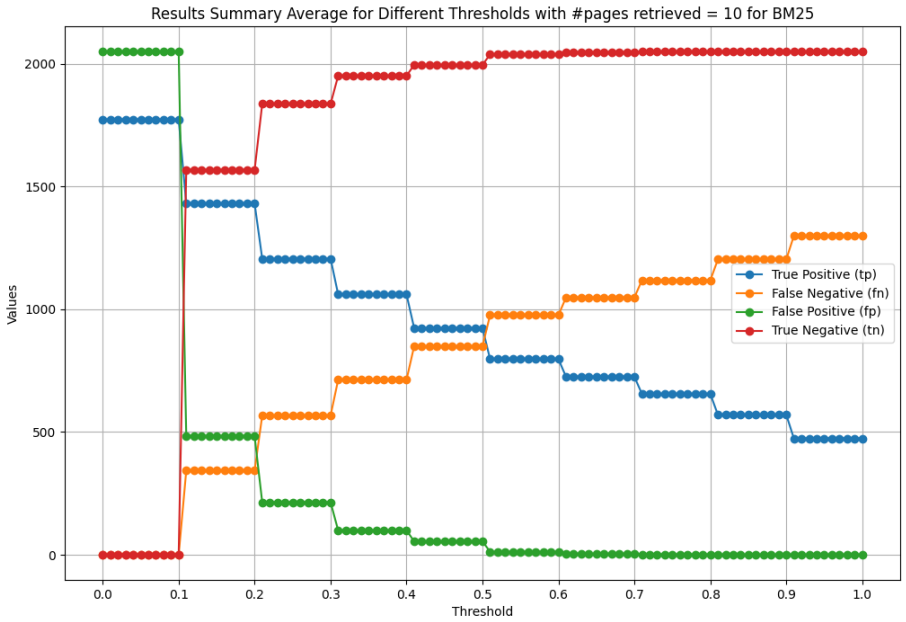

Figure 6.4: **Frequency** for **BM25**, for all ministries averaged, with n=10.

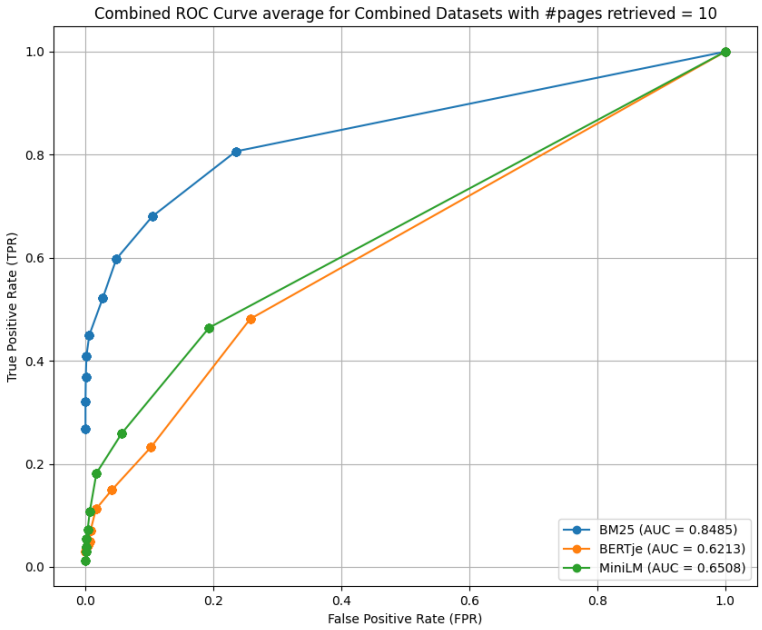

Figure 6.5: **ROC Frequency** for **BM25**, for all ministries averaged with n=10.

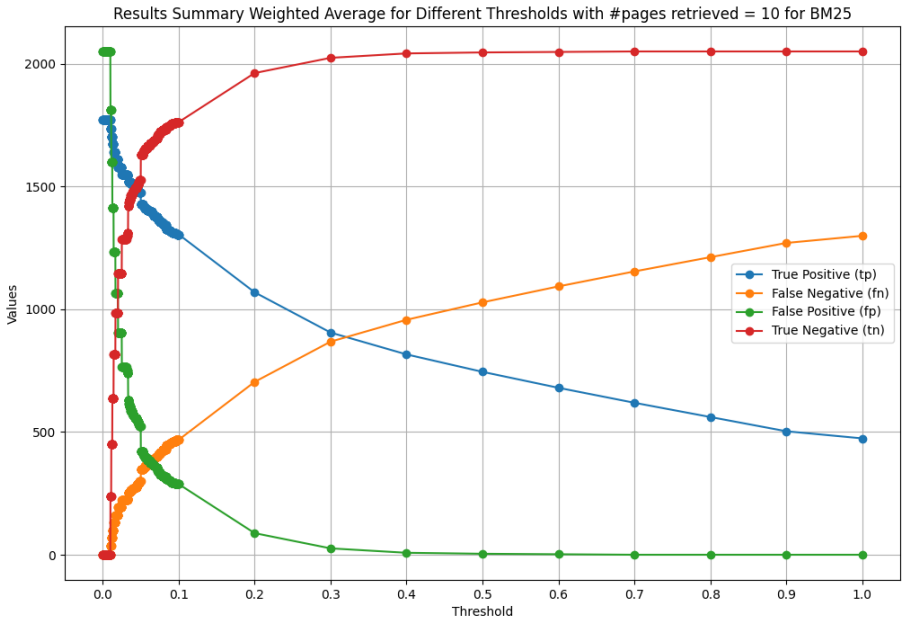

Figure 6.6: **Weighted Frequency** for **BM25**, for all ministries averaged with n=10.

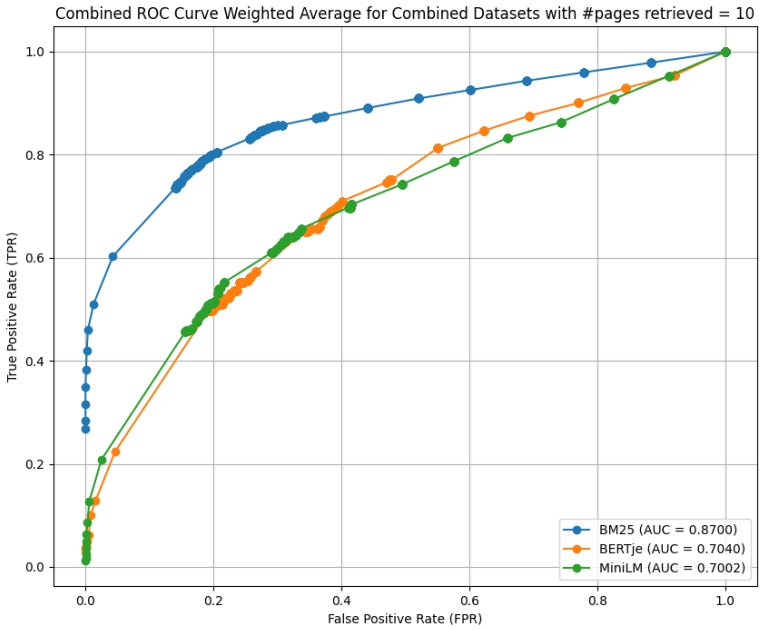

Figure 6.7: **ROC Weighted Frequency** for **the different models**, for all ministries averaged with n=10.

6.2.1. Time taken

The tables below present the amount of time the different processes have taken for all the ministries on average, which directly correlates to RQ3. The time taken for the ministries individually can be found in Appendix C.

In Table 6.4 we can see that preprocessing both the input files and the database does not take a large amount of time. Only generating the paraphrased request file and generating the keywords take a significant amount of time, since an LLM is used to generate these responses.

In Table 6.5, we can see the that greatest discrepancy in time is when creating the database. As for BM25, this was done in less than a second on average per dossier. However, for both dense retrieval methods, a single dossier could take up to multiple minutes. Creating embeddings is a heavy operational task for the GPU, hence the long operation time.

When querying the data, as can be seen in Table 6.6, we see that the time taken increases significantly with the size of the input file. We can also see that dense retrieval methods perform faster than BM25 on bigger input files. However, on smaller input files, BM25 performs faster than the dense retrieval methods.

Process	Time (sec)	Time per Dossier (sec)
Preprocessing Input		
Whole Request File	0	0
Generated Paraphrased Request File	4384.7	3.50
Generated Keywords based on Request File	3466.2	2.77
Real Words in Request File	3	0.00
Preprocessing Data		
All Documents	0	0
All Documents stopwords removed, stemmed	443.4	0.35
All Documents, real words	44.5	0.04

Table 6.4: Preprocessing Times

Process	Time (sec)	Time (hh:mm:ss)	Time per Dossier (sec)
Ingest BM25S			
All Documents	51.4		0.04102
All Documents stopwords removed, stemmed	43.48		0.03470
All Documents, real words	26.74		0.021
Ingest BERTje			
All Documents	722883	200:48:03	576.92
All Documents stopwords removed, stemmed	369181	102:33:01	294.64
All Documents, real words	346788	96:19:48	276.77
Ingest MiniLM			
All Documents	699662	194:21:02	558.39
All Documents stopwords removed, stemmed	362657	100:44:17	289.43
All Documents, real words	327050	90:50:50	261.01

Table 6.5: Database Creation Times for Different Methods

Process	Database	Time (sec)	Time per Dossier (sec)
Evaluatebm25			
Whole Request File	All Documents	409.08	0.00150
	All Documents no stopwords, stem	337.62	0.00124
	All Documents, real words	297.66	0.00109
(Generated) Paraphrased Request File	All Documents	3.21	0.00001
	All Documents no stopwords, stem	3.07	0.00001
	All Documents, real words	2.61	0.00001
(Generated) Keywords on Request File	All Documents	3.08	0.00001
	All Documents no stopwords, stem	2.90	0.00001
	All Documents, real words	2.48	0.00001
Real Words in Request File	All Documents	238.21	0.00088
	All Documents no stopwords, stem	194.82	0.00072
	All Documents, real words	216.11	0.00079
Evaluate BERTje			
Whole Request File	All Documents	336.16	0.00124
	All Documents no stopwords, stem	304.00	0.00112
	All Documents, real words	309.08	0.00114
(Generated) Paraphrased Request File	All Documents	40.42	0.00015
	All Documents no stopwords, stem	35.01	0.00013
	All Documents, real words	37.41	0.00014
(Generated) Keywords on Request File	All Documents	42.08	0.00015
	All Documents no stopwords, stem	34.84	0.00013
	All Documents, real words	36.84	0.00014
Real Words in Request File	All Documents	157.96	0.00058
	All Documents no stopwords, stem	156.82	0.00058
	All Documents, real words	155.59	0.00057
Evaluate MiniLM			
Whole Request File	All Documents	388.29	0.00143
	All Documents no stopwords, stem	361.04	0.00133
	All Documents, real words	330.11	0.00121
(Generated) Paraphrased Request File	All Documents	32.77	0.00012
	All Documents no stopwords, stem	27.88	0.00010
	All Documents, real words	29.39	0.00011
(Generated) Keywords on Request File	All Documents	30.50	0.00011
	All Documents no stopwords, stem	27.75	0.00010
	All Documents, real words	28.07	0.00010
Real Words in Request File	All Documents	206.87	0.00076
	All Documents no stopwords, stem	177.13	0.00065
	All Documents, real words	178.95	0.00066

Table 6.6: Evaluation Times for Different Models

7

Discussion

When assessing the effectiveness of different retrieval algorithms, as mentioned in RQ1, BM25 has consistently demonstrated superior performance compared to the other models across all evaluated metrics. Its ability to outperform dense retrieval models, sometimes by as much as a factor of ten, can indicate robustness and reliability of classic retrieval algorithms, even when considering modern models.

However, when analyzing results at the ministry level, it is hard to draw a conclusion due to its variability. Even the Ministry of Finance, which contains more numerical data than other ministries, shows similar trends to the other ministries. Generally, we observe that larger datasets tend to result in worse results, which is likely due to the increased noise in the retrieval process as dataset size grows. We can also observe that in the most of the cases in embeddings models, a combination of paraphrased query files with a non-preprocessed database works best, indicating the effectiveness of extracting the essence and reducing noise in a query file.

According to Massive Text Embedding Benchmark (MTEB) [35], embedding models do not consistently perform well across all tasks. While those optimized for similarity-based tasks excel in specific areas like semantic text similarity, their effectiveness can drop significantly when applied to specific tasks. The observed performance gap may be due to several factors:

- **Model Scalability:** As noted in the MTEB, embedding models' performance often scales with model size and the volume of training data. However, this improvement requires longer processing time and computational resources, which may make them less practical for large-scale applications. In our experiments, BM25s *simpler* scoring mechanism allows for more scalable and efficient handling of queries without the need for extensive fine-tuning.
- **Input data:** The input data across different experiments varies significantly in terms of scope and length. For instance, when using the raw dataset, queries span entire input files, often extending over multiple pages. However, only a small portion of these files typically contain the essence of the request, while the rest would be boilerplate text. This may challenge the embedding models to accurately capture the essential information. Conversely, in the paraphrased dataset, we attempt to isolate and capture the essence of these input files, which can be beneficial for embedding-based models. However, given the volume of data, paraphrasing is generated using a LLM, which in turn questions the performance of the LLM. Manual reviews reveal that some paraphrased queries capture the essence well, while others completely miss the essence of the information, which impacts the performance. Interestingly, even with paraphrased inputs, embeddings-based models perform competitively.
- **Data quality:** A notable challenge is the presence of incomplete or incorrect words within the dataset, which complicates semantic interpretation for embedding models. Embeddings

rely on contextual understanding, and inconsistencies within the input can lead to misinterpretations of semantic meaning. In contrast, BM25 relies on term frequency and inverse document frequency directly, making it less sensitive to minor inconsistencies or erroneous words, which could explain its stable performance even in noisier datasets.

We can observe notable differences in performance among the dense retrieval models. Although BERTje is trained specifically on Dutch data, and MiniLM is trained on English data, we still see that MiniLM has a slightly better performance across every metric. This performance gap likely comes from architectural differences between the models. BERTje, designed as a fill-mask model, is optimized for masked language tasks, which limits its capacity to capture semantic similarity between texts. MiniLM, in contrast, is specifically tuned for sentence similarity, enhancing its ability to interpret semantic connections between queries and documentseven when applied to Dutch data.

Additionally, the results indicate that query transformation can work for dense retrieval, while the raw query files always yield the best results in sparse retrieval. For the dense retrieval methods, we can see that a combination using the paraphrased query file often results into the best results, while, at the same time, using the keywords would result into the worst results. This is what we expected, since the paraphrased query was designed to improve the semantic understanding of the embedding models and the extracted keywords was an attempt to improve the results for sparse retrieval.

7.1. Weighted Re-evaluation

When looking at the weighted frequencies, we can see that we can successfully reduce the amount of documents that are not part of the real dossier. However, this reduction comes with a trade-off: some documents that are part of the dossier may be accidentally filtered out as well.

When examining the ROC curve, it reveals that the BM25 curve consistently lies above the dense embedding curves However, determining the optimal threshold remains hard to do, as there is no single ideal threshold. The best threshold would balance the true positive rate against the false positive rate, but this balance is hard to determine with our retrieval goals.

For example, when we take a threshold of 0.3 for Figure 6.6, we obtain the values:

- True Positive: 1550
- False Negative: 223
- False Positive: 766
- True Negative: 1284

This means the following:

- From all the 1773 results that were true (TP + FN), we filtered it down to only 1550 (TP) resulting in a loss of 223 pages (FP).
- From all the 2050 results that were false (FP + TN), we filtered it down to only 766 results that are false (FP), so we removed 1284 (TN) irrelevant pages.

While this approach proves effective in significantly reducing the volume of irrelevant documents a Woo coordinator must review, thereby significantly decreasing the processing time for a Woo request. It still raises an important question about the trade-off between efficiency and completeness. Although reducing the number of non-relevant documents can improve operational efficiency and decrease workload, the potential loss of valuable information may undermine the comprehensiveness of the dossier. With this, a critical consideration remains: is the gain in time saved worth the potential risk of excluding documents that could hold essential information?

To address the concern of losing valuable information, it is important to acknowledge that no dossier is ever guaranteed to contain all relevant data. Ensuring that as much relevant data is retrieved can be done by retrieving additional documents during the retrieval phase, though this would also increase the review burden. A manual review of every document is impractical and unrealistic. By using the weighted frequency approach while applying an appropriate threshold, we can effectively minimize the search space and retain essential information with more efficiency than a simple volume-based retrieval approach, despite the information loss.

It is also important to note that increasing the volume of retrieved documents, naturally increases the recall, but inversely impacts precision and MAP. As more documents are retrieved, we get a higher count of relevant documents, but at the same time, the retrieved documents also get a higher count of irrelevant documents.

7.2. Time

When evaluating the time taken to create the database, BM25 demonstrates a clear advantage in speed over dense retrieval methods. While BM25 requires only seconds to build a sparse database, embedding models need hours to generate and store dense vector representations, reflecting the substantial computational calculations associated with embeddings. This disparity highlights BM25 as a much more efficient option in terms of setup time, which could be beneficial in applications where rapid indexing is required.

This discrepancy is less apparent when querying the database. While BM25 is still a bit faster, both BM25 and embedding models exhibit comparable query response times, efficiently retrieving relevant documents in a matter of seconds. This similarity suggests that once indexed, the speed of accessing and processing queries is largely unaffected by the database type, making either method suitable for retrieval tasks.

With this finding, we can see BM25 as the more advantageous choice for our data set, not only in terms of setup time but also in achieving better retrieval performance.

7.3. Manual Checking

When analyzing retrieval frequencies within the Ministry of the Interior and Kingdom Relations dataset (241 queries) with n = 1 (retrieving only the top-ranked page per query), we observe distinct differences in how often individual pages are retrieved across models. Using BM25, the maximum retrieval frequency for any single page is 5. In contrast, MiniLM retrieves a single page up to 18 times, and BERTje retrieves one page as many as 150 times.

This discrepancy suggests that dense retrieval models like MiniLM and BERTje are more likely to repeatedly prioritize certain pages, possibly because the embeddings capture similarities that make specific pages appear highly relevant across multiple queries. BERTje, in particular, may show a strong retrieval bias due to its fill-mask architecture, which likely emphasizes frequently occurring terms or structures, which explains its performance compared to BM25.

In contrast, BM25's frequency-based scoring approach distributes retrievals more evenly across the dataset, indicating that it captures a broader relevance without over-prioritizing specific content. This suggests that while dense models may amplify certain pages based on embedding similarities, BM25 provides a more balanced retrieval, aligning well with the datasets diverse content.

8

Conclusion

In this thesis, we have identified one of the key reasons why Woo requests often face large delays. While there are multiple factors contributing to these delays, this thesis has focused on examining the retrieval mechanisms of the process. We identified flaws in current systems, and we are using state-of-the-art information retrieval techniques, to see if we can make this process better and faster.

With this research and with the experiments, we can answer the research questions.

RQ1: How effective are current state-of-the-art information retrieval methods in finding Woo requested information within Dutch Ministries?
Our research shows that dense retrieval methods, while promising in theory, currently perform poorly in comparison to BM25 when it comes to retrieving Woo requested information. Using the Woogle data for our experiments, we observed that dense retrieval models struggled to achieve the same level of results as BM25. This gap in performance can be attributed to various factors, including the nature of the data itself and the specificity of Woo requests, which have been discussed in Chapter 7.

However, it is essential to note that the scope of our experiment was limited, and the data did not perfectly replicate the full diversity of documents and queries processed within Dutch Ministries. This makes it difficult to definitively conclude the performance of these models in a real governmental application. dense retrieval, while under performing here, may show potential in different contexts where large-scale semantic understanding is critical. The BM25 model, on the other hand, consistently delivered reliable and relevant results, demonstrating both efficiency and high accuracy in handling the structured queries typical in governmental document retrieval.

The observed performance of BM25, combined with its quick processing capabilities, suggests it may be better suited for Woo request retrieval within Dutch Ministries. However, dense retrieval techniques should not be dismissed altogether, as their capability for nuanced contextual understanding may offer advantages. Over time, a combined approach could utilize BM25 with the semantic understanding of dense retrieval, adapting dynamically to the variety of requests handled by government document retrieval systems.

RQ2: What technological improvements can Dutch Ministries implement to enhance the accuracy and efficiency of document retrieval for Woo requests?
By adopting the algorithms and models discussed in this paper, Dutch Ministries can potentially improve their document retrieval process. They currently use boolean search together with a workflow to structurally construct their queries, which they can use with their Document Management System. While effective to a degree, this process is still quite tedious, as queries need to be con-

structed manually and are very prone to error. Oftentimes, a Woo coordinator has to refine their Woo query multiple times, before getting results that are acceptable to work with.

Transitioning from boolean search to a method like RAG would be a significant, ambitious, and perhaps an overly large step. Improving the discoverability of their documents with just an enhanced document retrieval system might already be a good initial step. Our experiments show that using state-of-the-art embedding-based models might not always be the optimal solution; while using state-of-the-art models is beneficial, it is also worth reminding ourselves that established methods like BM25 can, in some cases, offer better results. Such methods enhance search relevance and efficiency without the need for repeated query refinements like boolean search, offering a practical and effective alternative.

Although this paper does not present a direct comparison of the models to boolean search due to variations in processes and syntax, it does highlight their potential to reduce manual effort and improve search precision. Implementing these technologies could streamline Woo request processing, leading to faster retrieval times and more accurate results. Ministries could therefore benefit from integrating such state-of-the-art methods within their Document Management Systems, ultimately improving their ability to handle Woo requests faster, more effectively, and more reliably.

RQ3: How fast is it to create and retrieve information from databases using different methods for Woo requests, considering the performance of each method?
The estimated time required to create and retrieve information from databases for Woo requests varies significantly depending on the retrieval method. Dense retrieval methods, tend to be much slower in database creation and have shown mediocre performance with the available data. BM25 on the other hand, has a fast database creation and has also achieved better performance with the available data.

The time required for database creation is not the primary concern for users. While database creation time holds some importance, particularly for initial setup or large-scale updates, it is less critical because it is typically a one-time process (or periodic, when updates are necessary). What truly matters to users is the speed and efficiency of query processing, as rapid retrieval directly impacts their workflow and productivity.

In our analysis, we found no significant time difference in query speed between dense retrieval methods and BM25. Despite the slower database creation time associated with dense methods, both methods perform similarly in terms of retrieval speed, making them viable options for real-time querying. Therefore, while BM25 remains a consistently fast and reliable choice, dense methods can also be considered, especially in contexts where additional contextual insights might enhance retrieval outcomes without sacrificing query performance.

With this information, we hope that (Dutch) governmental institutions can gain valuable insights into improving their document retrieval processes. While the focus of this thesis was on Woo requests specifically, the knowledge gained can also be applied to other contexts within the government. An example of such a context is for retrieving parliamentary questions, where ministries must answer and provide data to the House of Representatives (in Dutch: Tweede Kamer der Staten-Generaal).

8.1. Future Work & Recommendations for SSC-ICT

While this thesis has identified effective approaches and highlighted current limitations in Woo request processing, there remain opportunities for SSC-ICT to refine and expand upon these findings to further improve document retrieval systems.

As mentioned in Section 1.3, SSC-ICT is actively researching and implementing a RAG system that has the potential to serve a wide variety of applications within the government. However, during initial observations, it was clear that foundational aspects of this system required more critical thinking. For example, when SSC-ICT began implementing the RAG application, embedding

models were chosen almost by default in the retrieval phase, without fully considering why certain retrieval methods performed poorly under specific conditions. In this thesis, we had a more critical approach. We explored options, including traditional algorithms like BM25, which, as shown in our analysis, can outperform embedding-based models given the nature of the current data. Adopting this balanced, evidence-based perspective can enable SSC-ICT to select the most effective tools for each specific use case, ensuring both accuracy and efficiency in retrieval processes.

In addition to a more rigorous evaluation of retrieval models, we propose a few strategies to enhance SSC-ICTs RAG implementation:

1. **Two-step retrieval pipeline**: A two-step approach could improve the retrieval process. In this setup, an initial query using BM25 could reduce the search space, retrieving a subset of relevant documents. Then a refined set would undergo a second round of retrieval using embeddings for deeper semantics. Essentially trying to take the best of both algorithms by taking the strengths of every approach.

2. **Hybrid retrieval approach**: Alternatively, a hybrid retrieval method could be used. In this setup, a weighted average between BM25 and embedding models can be used for the scoring system. This approach also attempts to take the best of both algorithms by taking them both into account.

3. **Data quality and consistency**: A critical factor of the performance in this paper (and any retrieval system process), is the quality of the data. Regardless of the retrieval model, poor data will lead to unreliable results. It remains essential for government institutions to prioritize consistent and standardized data storage. by investing in a better structure, SSC-ICT can ensure that the foundation of their retrieval system is more reliable, enhancing the overall performance and accuracy of Woo requests and other applications.

Looking ahead, SSC-ICT has the tools and insights to build a highly effective project. We believe that the findings from this research can support not only SSC-ICT but also the entire government across the board to enhance their retrieval processes. Setting a foundation for more efficient, accurate, and responsive information management in the public sector.

A

Evaluation of the Ministries - Full Results

This page is intentionally left blank.

Table A.1: **Mean Average Precision** at Different Amount of Pages Retrieved Values for **BERTje** on **Ministry of the Interior and Kingdom Relations**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.0416	0.0306	0.0243	0.0208	0.0184	0.0168	0.0158	0.0150	0.0145	0.0144
	stopwords	0.0024	0.0020	0.0018	0.0017	0.0015	0.0014	0.0013	0.0012	0.0011	0.0012
	real words	0.0173	0.0130	0.0109	0.0098	0.0087	0.0078	0.0071	0.0066	0.0062	0.0058
keywords	raw	0.0152	0.0107	0.0100	0.0090	0.0082	0.0074	0.0069	0.0065	0.0061	0.0064
	stopwords	0.0014	0.0016	0.0015	0.0017	0.0019	0.0018	0.0016	0.0014	0.0014	0.0013
	real words	0.0109	0.0084	0.0072	0.0064	0.0057	0.0054	0.0050	0.0047	0.0045	0.0043
paraphrase	raw	0.0513	0.0396	0.0328	0.0294	0.0265	0.0242	0.0223	0.0209	0.0204	0.0216
	stopwords	0.0033	0.0022	0.0017	0.0013	0.0011	0.0010	0.0009	0.0008	0.0007	0.0007
	real words	0.0216	0.0151	0.0124	0.0108	0.0096	0.0089	0.0083	0.0078	0.0075	0.0072
real words	raw	0.0191	0.0136	0.0105	0.0088	0.0077	0.0068	0.0061	0.0058	0.0057	0.0063
	stopwords	0.0065	0.0044	0.0031	0.0026	0.0022	0.0019	0.0018	0.0016	0.0015	0.0014
	real words	0.0333	0.0221	0.0176	0.0152	0.0134	0.0119	0.0110	0.0102	0.0096	0.0089

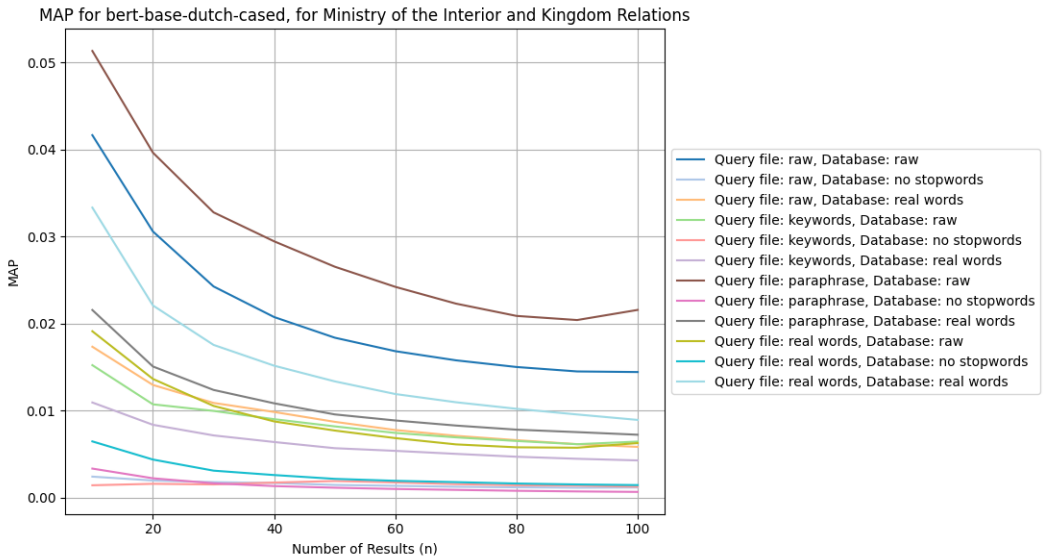

Figure A.1: Dataset: Ministry of the Interior and Kingdom Relations, Model: BERTje, Metric: MAP

Table A.2: **Precision** at Different Amount of Pages Retrieved for **BERTje** on **Ministry of the Interior and Kingdom Relations**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.0531	0.0423	0.0361	0.0335	0.0309	0.0287	0.0274	0.0255	0.0240	0.0221
	stopwords	0.0058	0.0068	0.0079	0.0084	0.0076	0.0075	0.0072	0.0072	0.0071	0.0076
	real words	0.0274	0.0239	0.0216	0.0206	0.0194	0.0183	0.0177	0.0168	0.0161	0.0159
keywords	raw	0.0261	0.0224	0.0227	0.0214	0.0205	0.0200	0.0194	0.0190	0.0184	0.0177
	stopwords	0.0050	0.0064	0.0065	0.0077	0.0083	0.0084	0.0076	0.0072	0.0071	0.0067
	real words	0.0203	0.0189	0.0177	0.0178	0.0164	0.0161	0.0154	0.0148	0.0147	0.0141
paraphrase	raw	0.0745	0.0672	0.0594	0.0563	0.0528	0.0501	0.0480	0.0459	0.0432	0.0394
	stopwords	0.0077	0.0066	0.0054	0.0049	0.0047	0.0045	0.0048	0.0046	0.0044	0.0044
	real words	0.0349	0.0291	0.0267	0.0260	0.0248	0.0238	0.0223	0.0213	0.0207	0.0206
real words	raw	0.0282	0.0228	0.0192	0.0174	0.0168	0.0163	0.0157	0.0157	0.0150	0.0140
	stopwords	0.0108	0.0087	0.0076	0.0072	0.0067	0.0066	0.0066	0.0063	0.0064	0.0065
	real words	0.0427	0.0349	0.0317	0.0293	0.0277	0.0255	0.0244	0.0235	0.0224	0.0215

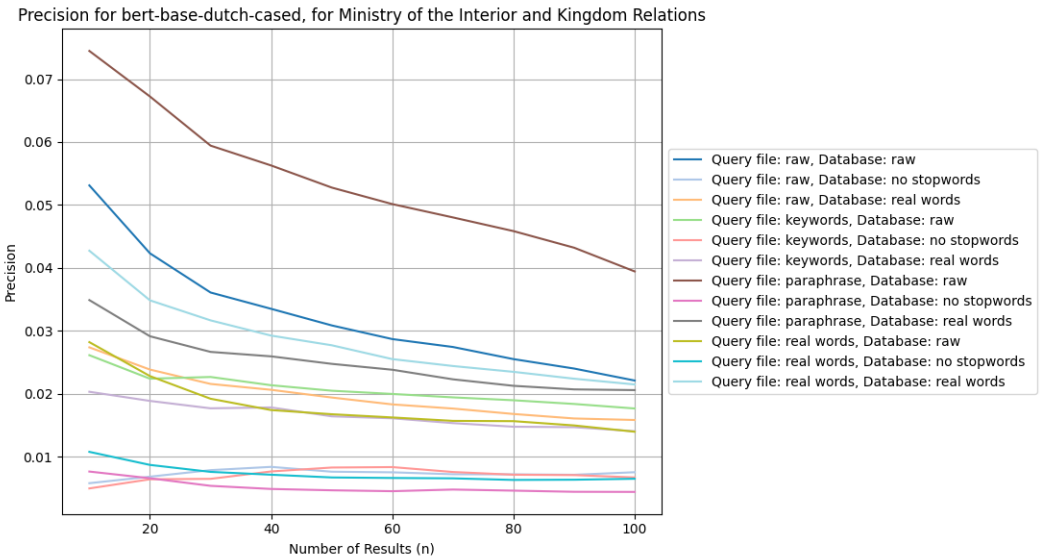

Figure A.2: Dataset: Ministry of the Interior and Kingdom Relations, Model: BERTje, Metric: Precision

Table A.3: **Recall** at Different Amount of Pages Retrieved for **BERTje** on **Ministry of the Interior and Kingdom Relations**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.0081	0.0107	0.0125	0.0137	0.0147	0.0157	0.0175	0.0183	0.0194	0.0204
	stopwords	0.0002	0.0003	0.0010	0.0013	0.0014	0.0022	0.0024	0.0025	0.0028	0.0038
	real words	0.0024	0.0035	0.0043	0.0053	0.0061	0.0079	0.0090	0.0103	0.0112	0.0130
keywords	raw	0.0017	0.0032	0.0044	0.0050	0.0055	0.0062	0.0071	0.0080	0.0083	0.0087
	stopwords	0.0007	0.0008	0.0010	0.0014	0.0017	0.0020	0.0020	0.0022	0.0023	0.0024
	real words	0.0008	0.0017	0.0028	0.0039	0.0043	0.0055	0.0058	0.0061	0.0068	0.0070
paraphrase	raw	0.0075	0.0133	0.0168	0.0196	0.0222	0.0259	0.0286	0.0305	0.0318	0.0322
	stopwords	0.0002	0.0003	0.0003	0.0006	0.0006	0.0009	0.0010	0.0012	0.0012	0.0013
	real words	0.0041	0.0051	0.0059	0.0074	0.0084	0.0095	0.0098	0.0106	0.0123	0.0134
real words	raw	0.0029	0.0043	0.0054	0.0065	0.0083	0.0110	0.0121	0.0137	0.0140	0.0146
	stopwords	0.0003	0.0004	0.0005	0.0006	0.0008	0.0013	0.0014	0.0019	0.0021	0.0024
	real words	0.0094	0.0130	0.0155	0.0175	0.0196	0.0203	0.0219	0.0235	0.0242	0.0251

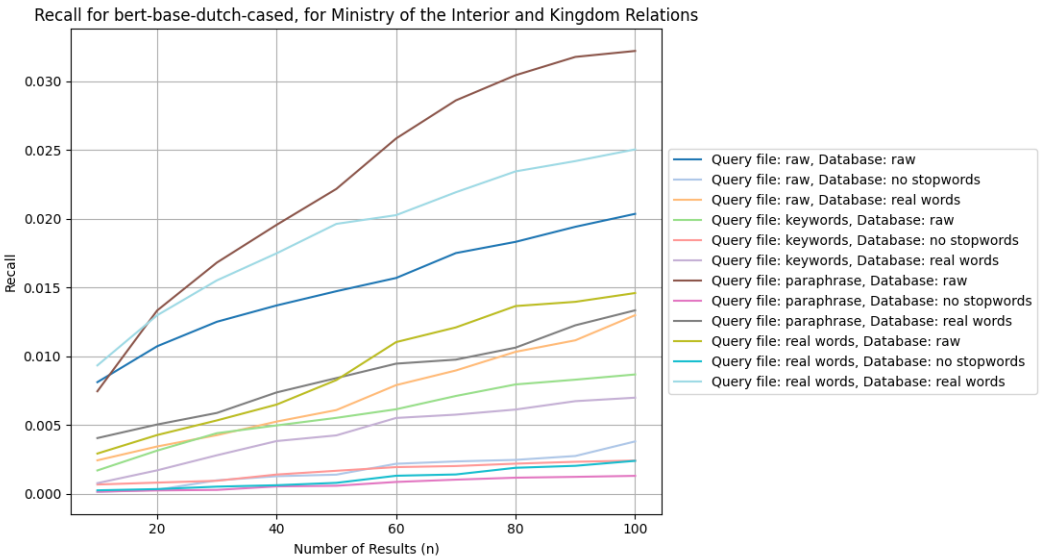

Figure A.3: Dataset: Ministry of the Interior and Kingdom Relations, Model: BERTje, Metric: Recall

Table A.4: **Mean Average Precision** at Different Amount of Pages Retrieved for **BM25** on **Ministry of the Interior and Kingdom Relations**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	**0.7268**	0.6599	0.6076	0.5712	0.5378	0.5121	0.4896	0.4698	0.4525	0.4361
	stopwords	0.7252	0.6551	0.6042	0.5683	0.5359	0.5095	0.4858	0.4652	0.4463	0.4282
	real words	0.6368	0.5774	0.5377	0.5065	0.4810	0.4573	0.4358	0.4168	0.3993	0.3826
keywords	raw	**0.1798**	0.1512	0.1323	0.1197	0.1101	0.1019	0.0957	0.0904	0.0856	0.0815
	stopwords	0.1344	0.1141	0.1018	0.0939	0.0881	0.0827	0.0779	0.0742	0.0713	0.0684
	real words	0.1283	0.1035	0.0894	0.0803	0.0730	0.0673	0.0625	0.0588	0.0553	0.0527
paraphrase	raw	**0.2846**	0.2327	0.2031	0.1832	0.1692	0.1580	0.1484	0.1410	0.1335	0.1269
	stopwords	0.2251	0.1865	0.1652	0.1501	0.1388	0.1294	0.1224	0.1154	0.1095	0.1046
	real words	0.2208	0.1798	0.1584	0.1440	0.1332	0.1239	0.1155	0.1091	0.1037	0.0988
real words	raw	**0.6148**	0.5596	0.5227	0.4948	0.4698	0.4455	0.4253	0.4055	0.3874	0.3703
	stopwords	0.6098	0.5496	0.5058	0.4725	0.4438	0.4185	0.3976	0.3778	0.3613	0.3453
	real words	0.6330	0.5740	0.5376	0.5092	0.4820	0.4595	0.4393	0.4200	0.4009	0.3838

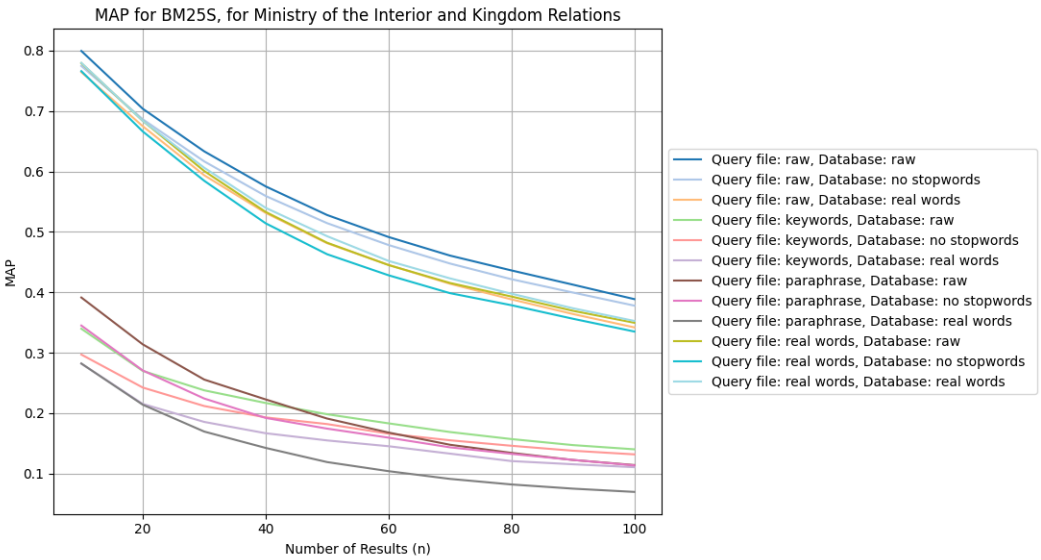

Figure A.4: Dataset: Ministry of the Interior and Kingdom Relations, Model: BM25, Metric: MAP

Table A.5: **Precision** at Different Amount of Pages Retrieved for **BM25** on **Ministry of the Interior and Kingdom Relations**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.7515	0.6890	0.6400	0.6067	0.5751	0.5503	0.5294	0.5109	0.4947	0.4787
	stopwords	0.7548	0.6880	0.6411	0.6072	0.5763	0.5505	0.5269	0.5073	0.4889	0.4710
	real words	0.6693	0.6170	0.5793	0.5488	0.5256	0.5023	0.4810	0.4636	0.4476	0.4316
keywords	raw	0.2154	0.1900	0.1715	0.1582	0.1475	0.1387	0.1315	0.1259	0.1198	0.1150
	stopwords	0.1598	0.1423	0.1296	0.1225	0.1165	0.1111	0.1056	0.1020	0.0988	0.0959
	real words	0.1593	0.1378	0.1245	0.1164	0.1084	0.1026	0.0974	0.0929	0.0882	0.0852
paraphrase	raw	0.3315	0.2877	0.2601	0.2390	0.2250	0.2135	0.2027	0.1948	0.1867	0.1790
	stopwords	0.2668	0.2351	0.2135	0.1987	0.1866	0.1767	0.1694	0.1616	0.1550	0.1494
	real words	0.2596	0.2221	0.2021	0.1888	0.1776	0.1681	0.1590	0.1520	0.1461	0.1406
real words	raw	0.6469	0.5963	0.5625	0.5366	0.5140	0.4898	0.4710	0.4531	0.4361	0.4195
	stopwords	0.6465	0.5896	0.5499	0.5179	0.4902	0.4658	0.4469	0.4279	0.4121	0.3965
	real words	0.6660	0.6110	0.5784	0.5517	0.5256	0.5035	0.4842	0.4662	0.4483	0.4320

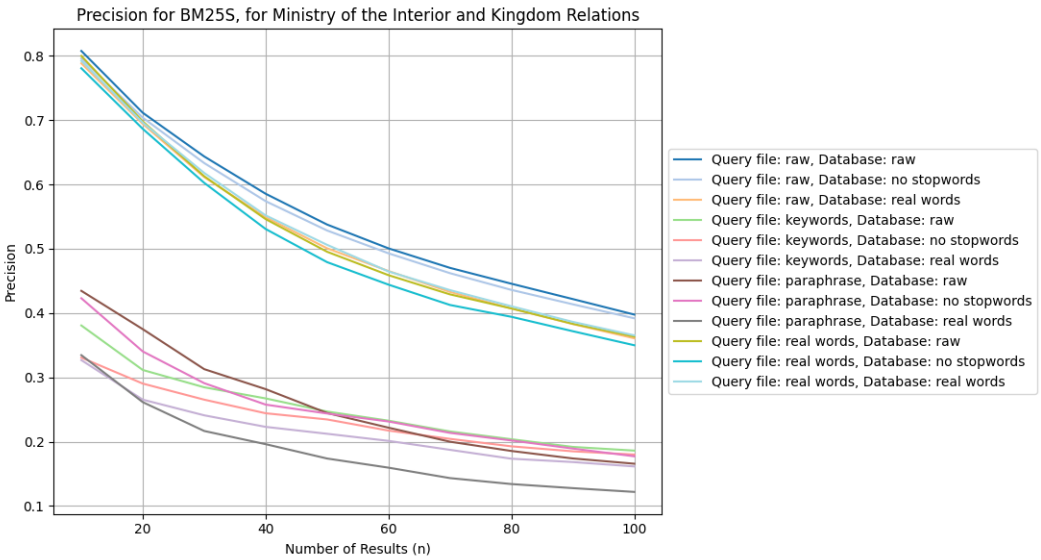

Figure A.5: Dataset: Ministry of the Interior and Kingdom Relations, Model: BM25, Metric: Precision

Table A.6: **Recall** at Different Amount of Pages Retrieved for **BM25** on **Ministry of the Interior and Kingdom Relations**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.1763	0.2509	0.2995	0.3408	0.3757	0.4004	0.4273	0.4558	0.4832	0.5024
	stopwords	0.1795	0.2540	0.3081	0.3467	0.3805	0.4105	0.4336	0.4569	0.4780	0.4989
	real words	0.1226	0.1935	0.2370	0.2687	0.3026	0.3257	0.3456	0.3685	0.3895	0.4083
keywords	raw	0.0436	0.0611	0.0739	0.0835	0.0931	0.1006	0.1059	0.1124	0.1176	0.1211
	stopwords	0.0291	0.0402	0.0487	0.0571	0.0634	0.0704	0.0750	0.0790	0.0841	0.0875
	real words	0.0241	0.0366	0.0474	0.0555	0.0617	0.0724	0.0775	0.0825	0.0873	0.0905
paraphrase	raw	0.0795	0.1044	0.1220	0.1345	0.1511	0.1611	0.1688	0.1777	0.1831	0.1876
	stopwords	0.0637	0.0843	0.0964	0.1109	0.1182	0.1246	0.1327	0.1380	0.1427	0.1482
	real words	0.0527	0.0699	0.0825	0.0934	0.1016	0.1126	0.1204	0.1254	0.1305	0.1364
real words	raw	0.1107	0.1730	0.2198	0.2515	0.2839	0.3062	0.3267	0.3486	0.3693	0.3876
	stopwords	0.1170	0.1748	0.2231	0.2533	0.2797	0.2988	0.3204	0.3400	0.3546	0.3678
	real words	0.1222	0.1904	0.2334	0.2692	0.3022	0.3266	0.3479	0.3709	0.3938	0.4120

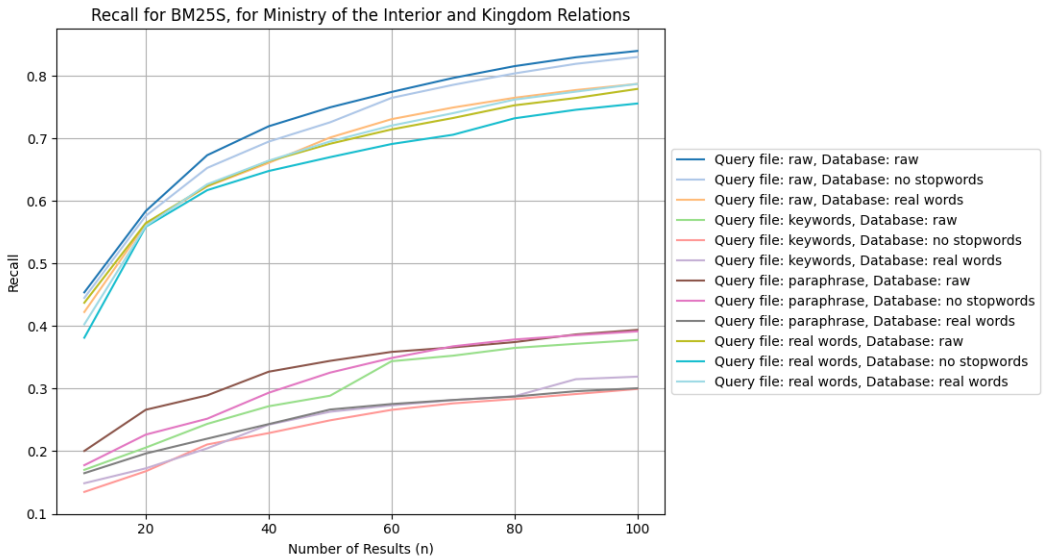

Figure A.6: Dataset: Ministry of the Interior and Kingdom Relations, Model: BM25, Metric: Recall

Table A.7: **Mean Average Precision** at Different Amount of Pages Retrieved for **MiniLM** on **Ministry of the Interior and Kingdom Relations**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.0743	0.0532	0.0427	0.0366	0.0329	0.0299	0.0274	0.0256	0.0244	0.0242
	stopwords	0.0425	0.0331	0.0278	0.0243	0.0218	0.0202	0.0190	0.0180	0.0171	0.0164
	real words	0.0399	0.0297	0.0235	0.0196	0.0171	0.0153	0.0138	0.0127	0.0117	0.0110
keywords	raw	0.0688	0.0494	0.0409	0.0365	0.0337	0.0315	0.0295	0.0283	0.0275	0.0277
	stopwords	0.0187	0.0136	0.0113	0.0103	0.0095	0.0090	0.0085	0.0082	0.0079	0.0077
	real words	0.0456	0.0360	0.0297	0.0268	0.0244	0.0226	0.0204	0.0189	0.0175	0.0164
paraphrase	raw	0.0649	0.0563	0.0492	0.0441	0.0402	0.0376	0.0354	0.0339	0.0331	0.0336
	stopwords	0.0257	0.0212	0.0177	0.0154	0.0141	0.0133	0.0126	0.0120	0.0115	0.0111
	real words	0.0415	0.0320	0.0271	0.0231	0.0209	0.0189	0.0174	0.0161	0.0150	0.0139
real words	raw	0.0270	0.0205	0.0169	0.0144	0.0127	0.0115	0.0105	0.0096	0.0089	0.0091
	stopwords	0.0091	0.0073	0.0054	0.0043	0.0037	0.0031	0.0029	0.0027	0.0025	0.0024
	real words	0.0448	0.0313	0.0234	0.0192	0.0161	0.0139	0.0124	0.0113	0.0102	0.0094

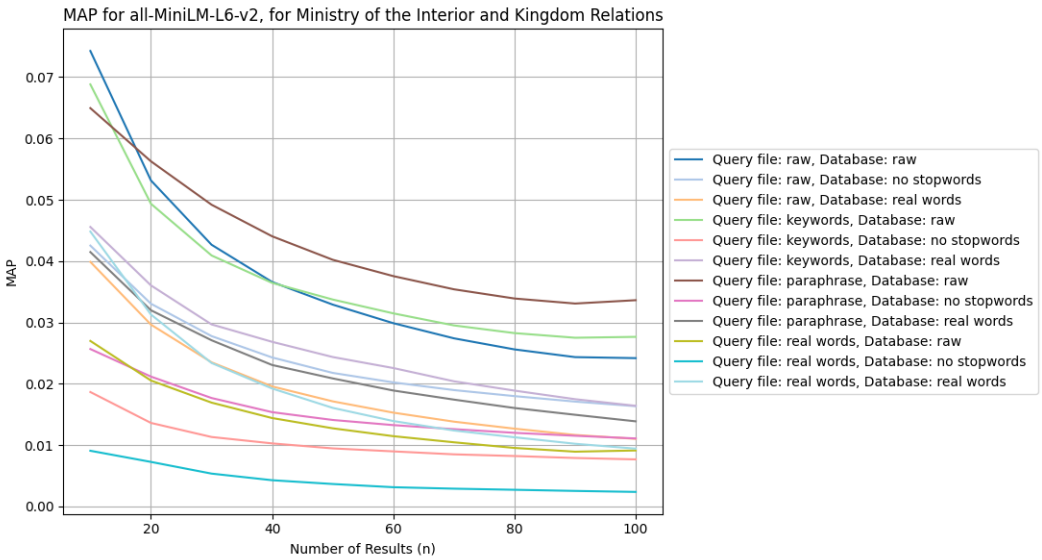

Figure A.7: Dataset: Ministry of the Interior and Kingdom Relations, Model: MiniLM, Metric: MAP

Table A.8: **Precision** at Different Amount of Pages Retrieved for **MiniLM** on **Ministry of the Interior and Kingdom Relations**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.0913	0.0751	0.0646	0.0590	0.0566	0.0530	0.0499	0.0474	0.0450	0.0421
	stopwords	0.0560	0.0467	0.0408	0.0376	0.0342	0.0323	0.0310	0.0303	0.0296	0.0285
	real words	0.0539	0.0481	0.0422	0.0381	0.0349	0.0332	0.0321	0.0307	0.0291	0.0284
keywords	raw	0.0876	0.0687	0.0610	0.0572	0.0545	0.0517	0.0491	0.0467	0.0438	0.0407
	stopwords	0.0282	0.0247	0.0216	0.0210	0.0207	0.0204	0.0197	0.0199	0.0195	0.0193
	real words	0.0581	0.0510	0.0454	0.0434	0.0410	0.0398	0.0372	0.0358	0.0343	0.0332
paraphrase	raw	0.0877	0.0840	0.0760	0.0706	0.0660	0.0625	0.0596	0.0575	0.0543	0.0508
	stopwords	0.0400	0.0379	0.0333	0.0305	0.0292	0.0287	0.0280	0.0272	0.0267	0.0262
	real words	0.0596	0.0521	0.0474	0.0436	0.0416	0.0395	0.0378	0.0357	0.0343	0.0329
real words	raw	0.0361	0.0336	0.0307	0.0282	0.0264	0.0250	0.0237	0.0224	0.0214	0.0203
	stopwords	0.0162	0.0170	0.0141	0.0128	0.0129	0.0121	0.0123	0.0123	0.0118	0.0116
	real words	0.0556	0.0429	0.0357	0.0320	0.0290	0.0268	0.0253	0.0241	0.0225	0.0217

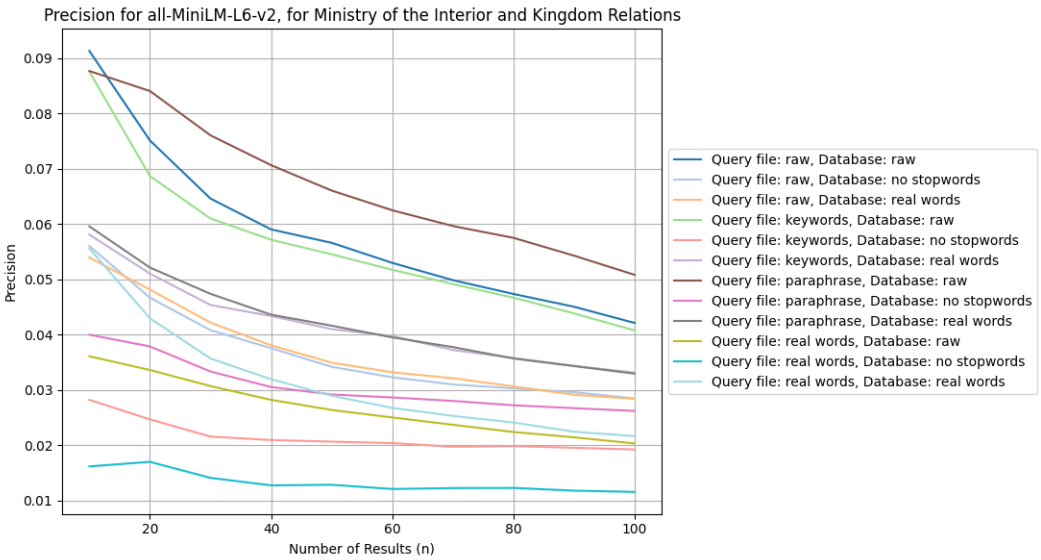

Figure A.8: Dataset: Ministry of the Interior and Kingdom Relations, Model: MiniLM, Metric: Precision

Table A.9: **Recall** at Different Amount of Pages Retrieved for **MiniLM** on **Ministry of the Interior and Kingdom Relations**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.0149	0.0205	0.0244	0.0262	0.0283	0.0298	0.0315	0.0329	0.0341	0.0353
	stopwords	0.0112	0.0128	0.0141	0.0152	0.0158	0.0165	0.0173	0.0182	0.0187	0.0199
	real words	0.0051	0.0082	0.0099	0.0124	0.0139	0.0150	0.0167	0.0176	0.0185	0.0197
keywords	raw	0.0101	0.0127	0.0149	0.0220	0.0241	0.0259	0.0273	0.0281	0.0302	0.0306
	stopwords	0.0040	0.0054	0.0078	0.0100	0.0106	0.0123	0.0132	0.0205	0.0215	0.0222
	real words	0.0030	0.0061	0.0082	0.0112	0.0124	0.0137	0.0146	0.0177	0.0192	0.0198
paraphrase	raw	0.0202	0.0253	0.0292	0.0322	0.0342	0.0358	0.0370	0.0390	0.0412	0.0422
	stopwords	0.0021	0.0124	0.0151	0.0161	0.0191	0.0203	0.0209	0.0222	0.0233	0.0239
	real words	0.0149	0.0186	0.0206	0.0223	0.0241	0.0259	0.0277	0.0289	0.0304	0.0326
real words	raw	0.0038	0.0057	0.0079	0.0089	0.0105	0.0115	0.0130	0.0140	0.0152	0.0155
	stopwords	0.0016	0.0067	0.0081	0.0084	0.0095	0.0098	0.0104	0.0111	0.0119	0.0125
	real words	0.0098	0.0120	0.0136	0.0160	0.0172	0.0182	0.0195	0.0207	0.0212	0.0221

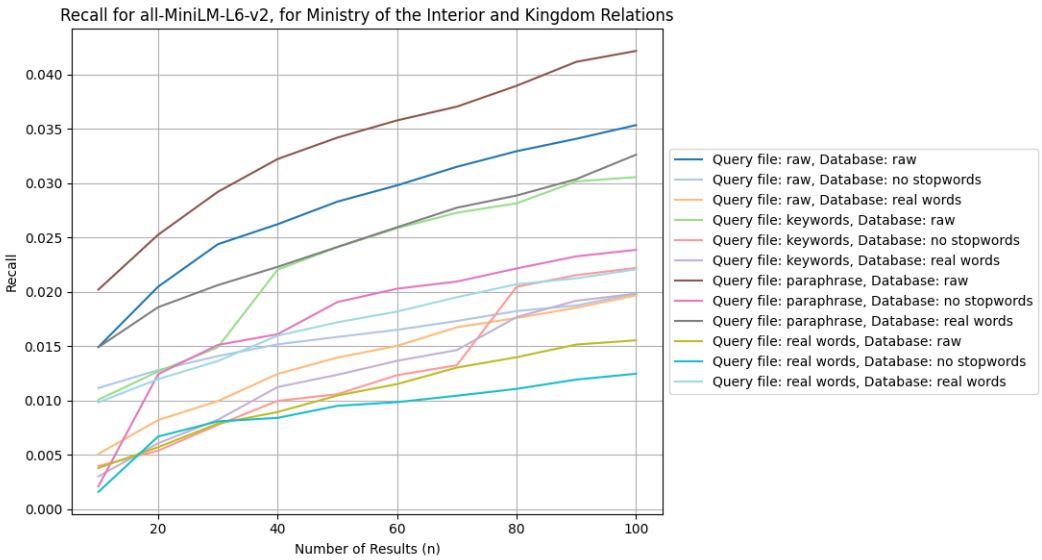

Figure A.9: Dataset: Ministry of the Interior and Kingdom Relations, Model: MiniLM, Metric: Recall

Table A.10: **Mean Average Precision** at Different Amount of Pages Retrieved Values for **BERTje** on **Ministry of General Affairs**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.1665	0.1191	0.0986	0.0855	0.0736	0.0649	0.0571	0.0528	0.0527	0.0558
	stopwords	0.0563	0.0525	0.0424	0.0348	0.0290	0.0264	0.0243	0.0228	0.0212	0.0202
	real words	0.0619	0.0460	0.0339	0.0307	0.0284	0.0266	0.0236	0.0219	0.0215	0.0202
keywords	raw	0.0637	0.0544	0.0448	0.0373	0.0335	0.0302	0.0290	0.0279	0.0283	0.0316
	stopwords	0.0316	0.0244	0.0190	0.0156	0.0134	0.0120	0.0113	0.0107	0.0102	0.0097
	real words	0.0526	0.0440	0.0399	0.0393	0.0346	0.0323	0.0318	0.0312	0.0289	0.0273
paraphrase	raw	0.1090	0.0714	0.0581	0.0492	0.0432	0.0390	0.0364	0.0361	0.0377	0.0412
	stopwords	0.0420	0.0294	0.0228	0.0204	0.0179	0.0157	0.0144	0.0132	0.0124	0.0116
	real words	0.0747	0.0475	0.0368	0.0323	0.0289	0.0272	0.0250	0.0237	0.0221	0.0219
real words	raw	0.0949	0.0696	0.0582	0.0477	0.0429	0.0386	0.0356	0.0348	0.0366	0.0394
	stopwords	0.0575	0.0426	0.0385	0.0308	0.0253	0.0219	0.0199	0.0187	0.0170	0.0159
	real words	0.1069	0.0765	0.0631	0.0527	0.0462	0.0403	0.0351	0.0329	0.0307	0.0291

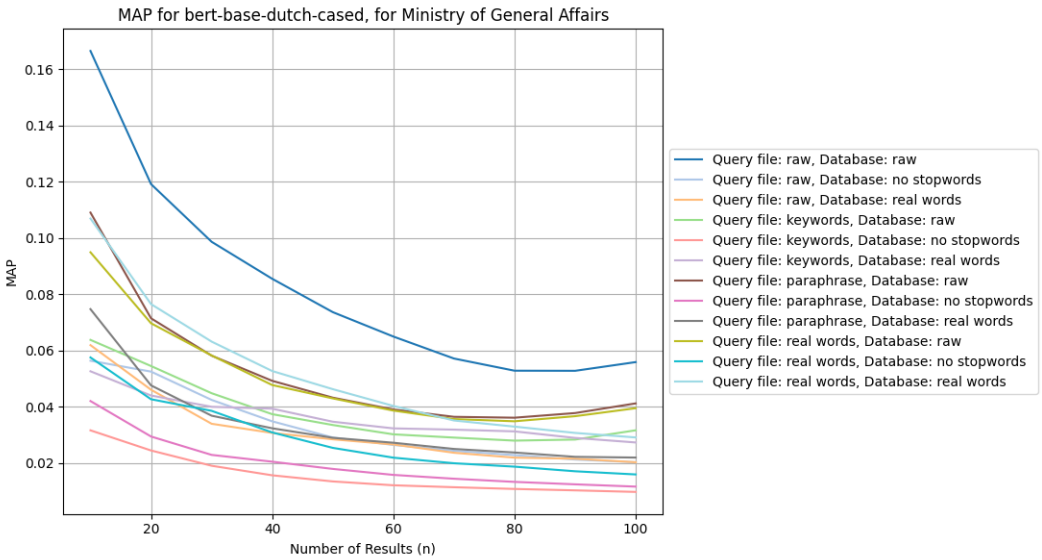

Figure A.10: Dataset: Ministry of General Affairs, Model: BERTje, Metric: MAP

Table A.11: **Precision** at Different Amount of Pages Retrieved Values for **BERTje** on **Ministry of General Affairs**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.2038	0.1596	0.1474	0.1317	0.1215	0.1122	0.1044	0.1000	0.0919	0.0827
	stopwords	0.0846	0.0808	0.0705	0.0654	0.0608	0.0622	0.0588	0.0582	0.0564	0.0573
	real words	0.1000	0.0923	0.0782	0.0837	0.0831	0.0782	0.0709	0.0678	0.0697	0.0673
keywords	raw	0.1000	0.1096	0.1013	0.0875	0.0838	0.0788	0.0802	0.0764	0.0744	0.0685
	stopwords	0.0500	0.0500	0.0487	0.0462	0.0454	0.0449	0.0478	0.0471	0.0487	0.0473
	real words	0.1077	0.0923	0.0885	0.0894	0.0815	0.0769	0.0775	0.0788	0.0744	0.0719
paraphrase	raw	0.1692	0.1231	0.1128	0.1019	0.0946	0.0917	0.0868	0.0827	0.0765	0.0688
	stopwords	0.0615	0.0538	0.0564	0.0567	0.0531	0.0500	0.0505	0.0510	0.0500	0.0481
	real words	0.1269	0.0962	0.0846	0.0837	0.0777	0.0776	0.0731	0.0702	0.0667	0.0677
real words	raw	0.1500	0.1250	0.1103	0.0981	0.0915	0.0853	0.0846	0.0837	0.0774	0.0696
	stopwords	0.0808	0.0750	0.0718	0.0635	0.0569	0.0519	0.0522	0.0529	0.0509	0.0500
	real words	0.1423	0.1154	0.1128	0.1010	0.0931	0.0865	0.0775	0.0769	0.0748	0.0727

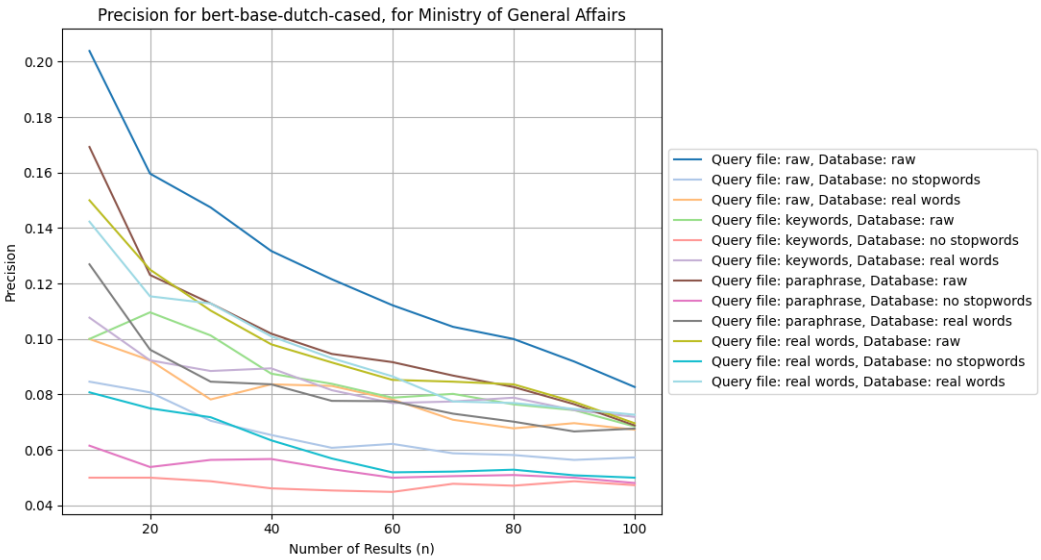

Figure A.11: Dataset: Ministry of General Affairs, Model: BERTje, Metric: Precision

Table A.12: **Recall** at Different Amount of Pages Retrieved Values for **BERTje** on **Ministry of General Affairs**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.0674	0.1081	0.1442	0.1546	0.1670	0.1735	0.1880	0.2019	0.2044	0.2044
	stopwords	0.0135	0.0223	0.0263	0.0336	0.0419	0.0501	0.0527	0.0779	0.0818	0.1066
	real words	0.0692	0.0823	0.1052	0.1287	0.1376	0.1450	0.1467	0.1522	0.1677	0.1725
keywords	raw	0.0721	0.0907	0.1045	0.1081	0.1153	0.1201	0.1335	0.1378	0.1421	0.1444
	stopwords	0.0057	0.0102	0.0227	0.0275	0.0332	0.0391	0.0476	0.0513	0.0651	0.0743
	real words	0.0699	0.0795	0.0868	0.0961	0.1042	0.1100	0.1173	0.1263	0.1297	0.1364
paraphrase	raw	0.0979	0.1064	0.1224	0.1310	0.1548	0.1631	0.1690	0.1723	0.1735	0.1735
	stopwords	0.0075	0.0133	0.0229	0.0288	0.0341	0.0367	0.0425	0.0576	0.0657	0.0745
	real words	0.0829	0.0886	0.1207	0.1294	0.1341	0.1568	0.1643	0.1686	0.1731	0.1817
real words	raw	0.0597	0.0812	0.1078	0.1166	0.1242	0.1326	0.1487	0.1551	0.1759	0.1759
	stopwords	0.0160	0.0248	0.0301	0.0333	0.0364	0.0408	0.0488	0.0568	0.0641	0.0682
	real words	0.0629	0.0763	0.1026	0.1180	0.1349	0.1455	0.1497	0.1567	0.1616	0.1740

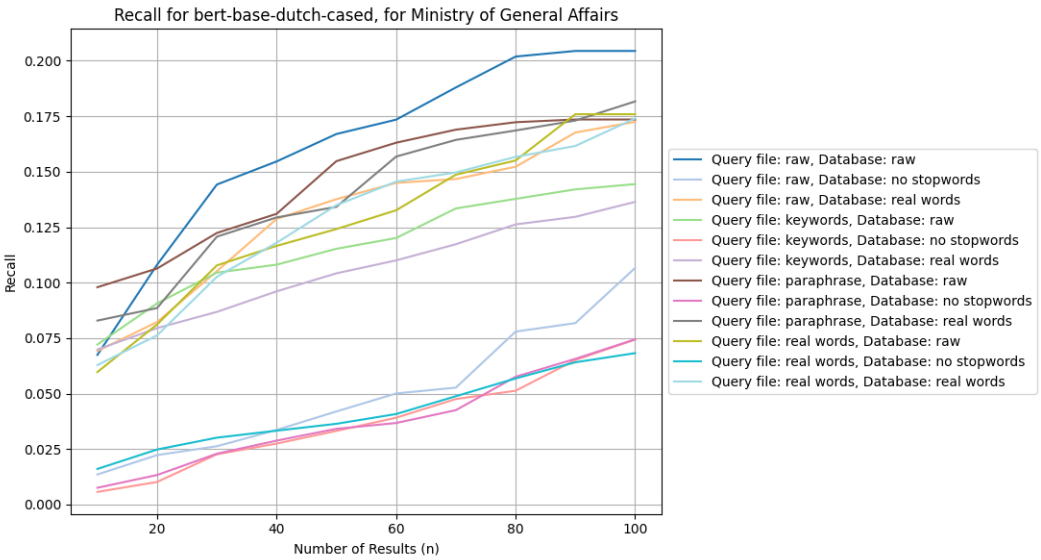

Figure A.12: Dataset: Ministry of General Affairs, Model: BERTje, Metric: Recall

Table A.13: **Mean Average Precision** at Different Amount of Pages Retrieved **Values** for BM25 on **Ministry of General Affairs**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.7994	0.7037	0.6332	0.5753	0.5279	0.4915	0.4608	0.4362	0.4127	0.3887
	stopwords	0.7746	0.6863	0.6166	0.5594	0.5144	0.4783	0.4476	0.4217	0.3996	0.3780
	real words	0.7643	0.6748	0.5937	0.5314	0.4813	0.4453	0.4142	0.3883	0.3641	0.3420
keywords	raw	0.3398	0.2699	0.2378	0.2169	0.1983	0.1832	0.1689	0.1572	0.1472	0.1403
	stopwords	0.2970	0.2424	0.2119	0.1931	0.1820	0.1661	0.1552	0.1461	0.1379	0.1319
	real words	0.2820	0.2153	0.1856	0.1668	0.1549	0.1454	0.1331	0.1208	0.1155	0.1107
paraphrase	raw	0.3914	0.3139	0.2556	0.2228	0.1911	0.1680	0.1478	0.1344	0.1226	0.1142
	stopwords	0.3450	0.2707	0.2243	0.1922	0.1745	0.1595	0.1435	0.1327	0.1227	0.1137
	real words	0.2822	0.2140	0.1696	0.1426	0.1191	0.1041	0.0913	0.0821	0.0752	0.0698
real words	raw	0.7791	0.6837	0.6007	0.5331	0.4821	0.4451	0.4155	0.3930	0.3694	0.3496
	stopwords	0.7660	0.6662	0.5846	0.5142	0.4630	0.4282	0.3986	0.3787	0.3563	0.3351
	real words	0.7796	0.6837	0.6061	0.5397	0.4928	0.4519	0.4231	0.3977	0.3736	0.3527

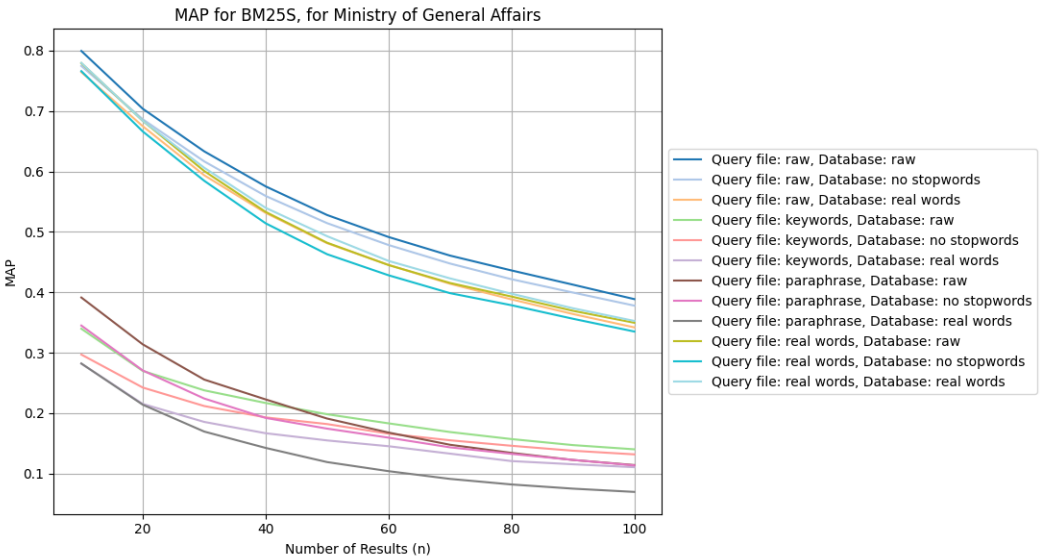

Figure A.13: Dataset: Ministry of General Affairs, Model: BM25, Metric: MAP

Table A.14: **Precision** at Different Amount of Pages Retrieved Values for **BM25** on **Ministry of General Affairs**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.8077	0.7115	0.6436	0.5856	0.5377	0.5006	0.4703	0.4457	0.4218	0.3977
	stopwords	0.7962	0.7038	0.6333	0.5740	0.5285	0.4929	0.4621	0.4361	0.4137	0.3919
	real words	0.7885	0.6942	0.6115	0.5490	0.5008	0.4654	0.4330	0.4072	0.3829	0.3608
keywords	raw	0.3808	0.3115	0.2846	0.2673	0.2469	0.2327	0.2159	0.2038	0.1919	0.1862
	stopwords	0.3308	0.2904	0.2654	0.2442	0.2346	0.2173	0.2044	0.1928	0.1850	0.1800
	real words	0.3269	0.2654	0.2410	0.2231	0.2123	0.2013	0.1874	0.1736	0.1684	0.1615
paraphrase	raw	0.4346	0.3750	0.3128	0.2817	0.2446	0.2218	0.2000	0.1856	0.1739	0.1658
	stopwords	0.4231	0.3404	0.2910	0.2577	0.2438	0.2314	0.2137	0.2019	0.1893	0.1773
	real words	0.3346	0.2615	0.2167	0.1962	0.1738	0.1596	0.1434	0.1341	0.1278	0.1219
real words	raw	0.8000	0.6981	0.6128	0.5462	0.4954	0.4590	0.4291	0.4072	0.3833	0.3635
	stopwords	0.7808	0.6865	0.6026	0.5308	0.4792	0.4442	0.4126	0.3942	0.3718	0.3500
	real words	0.7923	0.6962	0.6179	0.5519	0.5062	0.4647	0.4357	0.4106	0.3863	0.3654

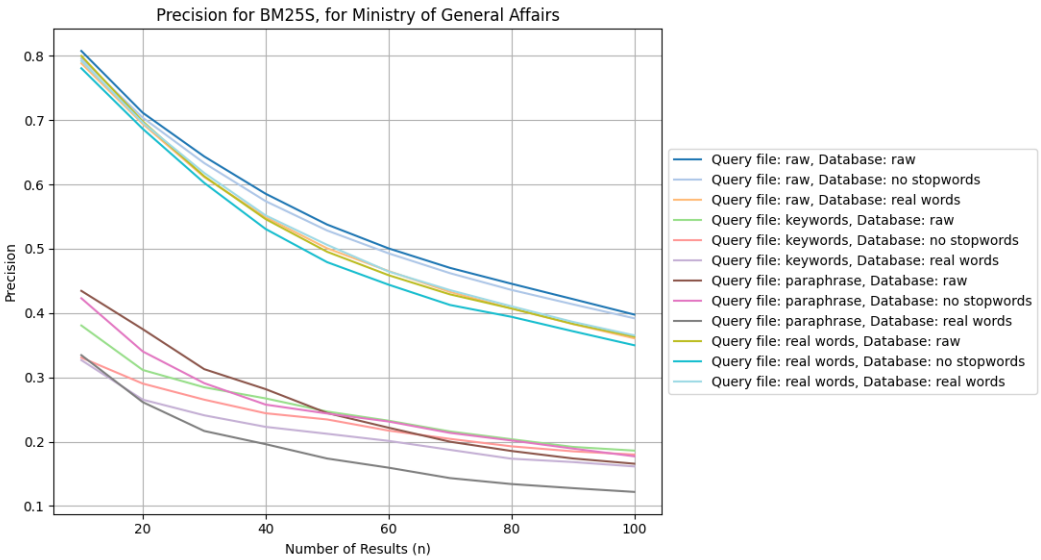

Figure A.14: Dataset: Ministry of General Affairs, Model: BM25, Metric: Precision

Table A.15: **Recall** at Different Amount of Pages Retrieved Values for **BM25** on **Ministry of General Affairs**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.4538	0.5838	0.6731	0.7193	0.7496	0.7742	0.7965	0.8154	0.8296	0.8397
	stopwords	0.4450	0.5761	0.6528	0.6948	0.7257	0.7647	0.7855	0.8038	0.8191	0.8300
	real words	0.4225	0.5633	0.6233	0.6608	0.7013	0.7307	0.7493	0.7648	0.7772	0.7871
keywords	raw	0.1702	0.2058	0.2437	0.2718	0.2887	0.3438	0.3527	0.3651	0.3716	0.3777
	stopwords	0.1349	0.1679	0.2108	0.2289	0.2494	0.2661	0.2764	0.2834	0.2914	0.2995
	real words	0.1488	0.1726	0.2045	0.2424	0.2631	0.2733	0.2816	0.2878	0.3151	0.3192
paraphrase	raw	0.2004	0.2662	0.2893	0.3270	0.3444	0.3587	0.3659	0.3744	0.3866	0.3939
	stopwords	0.1776	0.2265	0.2519	0.2934	0.3256	0.3490	0.3677	0.3785	0.3854	0.3915
	real words	0.1649	0.1963	0.2200	0.2433	0.2666	0.2754	0.2819	0.2874	0.2959	0.3006
real words	raw	0.4373	0.5646	0.6240	0.6635	0.6914	0.7143	0.7325	0.7528	0.7646	0.7790
	stopwords	0.3814	0.5588	0.6172	0.6478	0.6700	0.6910	0.7058	0.7320	0.7457	0.7557
	real words	0.4028	0.5599	0.6266	0.6643	0.6951	0.7204	0.7403	0.7619	0.7747	0.7870

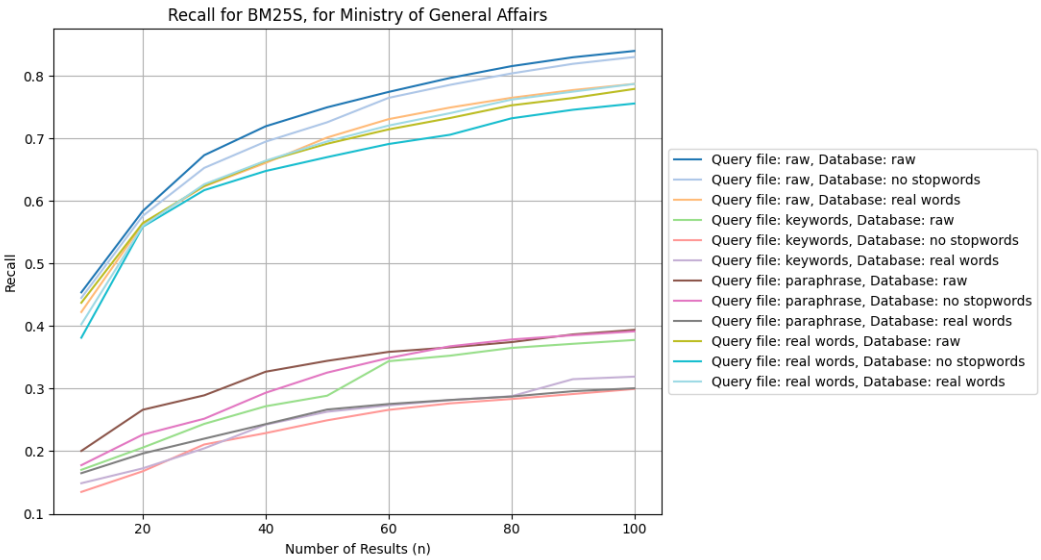

Figure A.15: Dataset: Ministry of General Affairs, Model: BM25, Metric: Recall

Table A.16: **Mean Average Precision** at Different Amount of Pages Retrieved Values for **MiniLM** on **Ministry of General Affairs**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.2102	0.1627	0.1317	0.1120	0.0994	0.0896	0.0800	0.0734	0.0706	0.0717
	stopwords	0.1238	0.0786	0.0642	0.0548	0.0483	0.0434	0.0408	0.0388	0.0376	0.0361
	real words	0.1178	0.0886	0.0802	0.0736	0.0709	0.0647	0.0605	0.0574	0.0542	0.0516
keywords	raw	0.2427	0.1798	0.1610	0.1435	0.1343	0.1282	0.1222	0.1177	0.1150	0.1164
	stopwords	0.1882	0.1367	0.1117	0.0959	0.0855	0.0776	0.0720	0.0663	0.0611	0.0575
	real words	0.1860	0.1423	0.1232	0.1140	0.1049	0.0972	0.0930	0.0892	0.0848	0.0798
paraphrase	raw	0.1982	0.1501	0.1275	0.1142	0.1045	0.0972	0.0920	0.0894	0.0873	0.0896
	stopwords	0.1439	0.1162	0.0958	0.0870	0.0796	0.0726	0.0678	0.0622	0.0584	0.0553
	real words	0.1339	0.1052	0.0967	0.0886	0.0854	0.0812	0.0785	0.0745	0.0717	0.0691
real words	raw	0.1724	0.1242	0.0988	0.0886	0.0807	0.0746	0.0707	0.0677	0.0664	0.0676
	stopwords	0.1286	0.0894	0.0726	0.0625	0.0554	0.0505	0.0460	0.0433	0.0404	0.0381
	real words	0.1754	0.1235	0.1003	0.0892	0.0811	0.0751	0.0691	0.0668	0.0630	0.0599

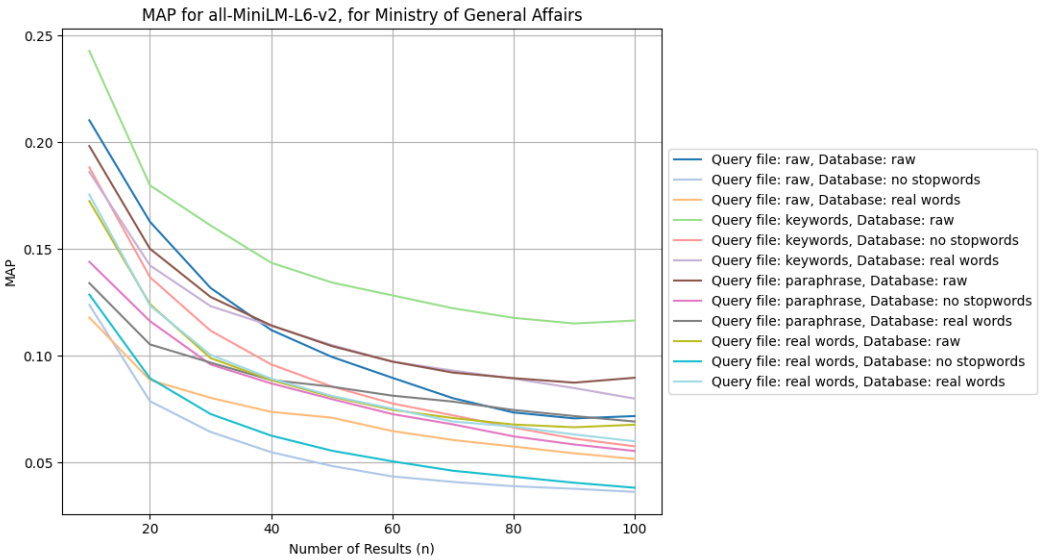

Figure A.16: Dataset: Ministry of General Affairs, Model: MiniLM, Metric: MAP

Table A.17: **Precision** at Different Amount of Pages Retrieved Values for **MiniLM** on **Ministry of General Affairs**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.2538	0.2038	0.1705	0.1519	0.1423	0.1301	0.1203	0.1130	0.1034	0.0946
	stopwords	0.1615	0.1250	0.1128	0.1038	0.0992	0.0936	0.0901	0.0899	0.0885	0.0873
	real words	0.1577	0.1346	0.1295	0.1192	0.1154	0.1083	0.1027	0.0990	0.0957	0.0938
keywords	raw	0.2577	0.2077	0.2000	0.1827	0.1738	0.1686	0.1610	0.1553	0.1432	0.1296
	stopwords	0.2385	0.1846	0.1577	0.1413	0.1331	0.1263	0.1203	0.1130	0.1064	0.1027
	real words	0.2269	0.1788	0.1628	0.1606	0.1500	0.1417	0.1390	0.1361	0.1299	0.1242
paraphrase	raw	0.2577	0.2135	0.1910	0.1740	0.1600	0.1538	0.1456	0.1385	0.1295	0.1173
	stopwords	0.1885	0.1615	0.1372	0.1308	0.1215	0.1147	0.1104	0.1067	0.1043	0.1000
	real words	0.1846	0.1615	0.1615	0.1481	0.1462	0.1391	0.1352	0.1303	0.1239	0.1196
real words	raw	0.2077	0.1673	0.1372	0.1327	0.1223	0.1128	0.1088	0.1038	0.0966	0.0877
	stopwords	0.1692	0.1365	0.1205	0.1087	0.1054	0.0968	0.0901	0.0856	0.0812	0.0792
	real words	0.2115	0.1654	0.1385	0.1260	0.1185	0.1109	0.1049	0.1038	0.1000	0.0962

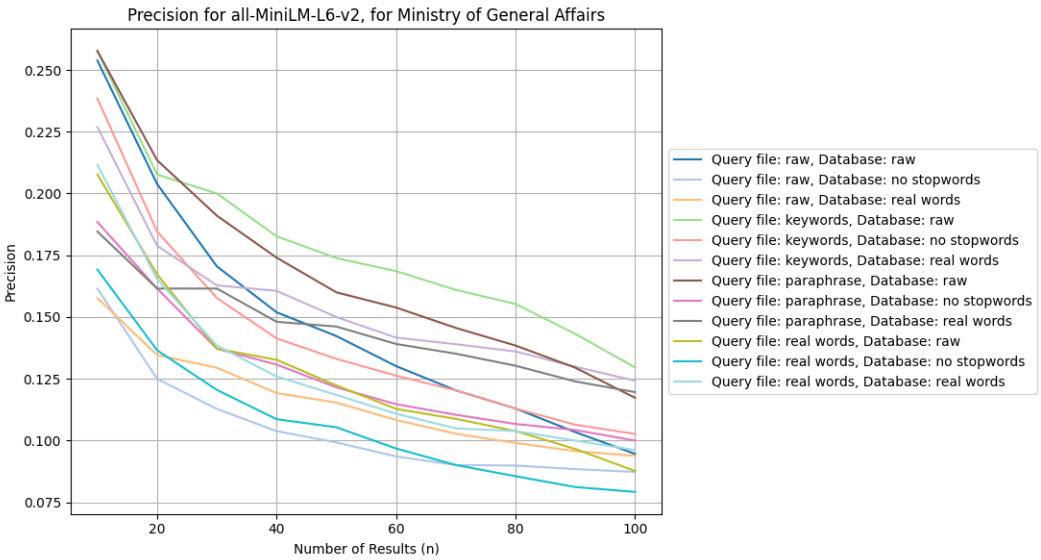

Figure A.17: Dataset: Ministry of General Affairs, Model: MiniLM, Metric: Precision

Table A.18: **Recall** at Different Amount of Pages Retrieved Values for **MiniLM** on **Ministry of General Affairs**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.1137	0.1278	0.1507	0.2085	0.2194	0.2237	0.2374	0.2486	0.2540	0.2548
	stopwords	0.0725	0.1056	0.1190	0.1275	0.1584	0.1727	0.1795	0.1881	0.1990	0.2079
	real words	0.0877	0.0960	0.1104	0.1204	0.1347	0.1397	0.1453	0.1489	0.1537	0.1880
keywords	raw	0.0950	0.1112	0.1307	0.1432	0.1538	0.1630	0.1714	0.1781	0.1810	0.1813
	stopwords	0.0543	0.0682	0.0755	0.0916	0.1033	0.1111	0.1185	0.1231	0.1260	0.1323
	real words	0.0696	0.0848	0.0986	0.1322	0.1413	0.1475	0.1739	0.1831	0.1906	0.1958
paraphrase	raw	0.1322	0.1789	0.1990	0.2143	0.2256	0.2380	0.2444	0.2511	0.2576	0.2579
	stopwords	0.0335	0.0473	0.0604	0.0728	0.0974	0.1232	0.1327	0.1447	0.1660	0.1700
	real words	0.0843	0.1062	0.1220	0.1296	0.1524	0.1616	0.1691	0.1794	0.1821	0.1872
real words	raw	0.0575	0.0935	0.1033	0.1381	0.1415	0.1470	0.1598	0.1834	0.1912	0.1914
	stopwords	0.0420	0.0675	0.0958	0.1020	0.1151	0.1185	0.1233	0.1277	0.1367	0.1413
	real words	0.0669	0.0938	0.1078	0.1128	0.1239	0.1320	0.1507	0.1597	0.1676	0.1709

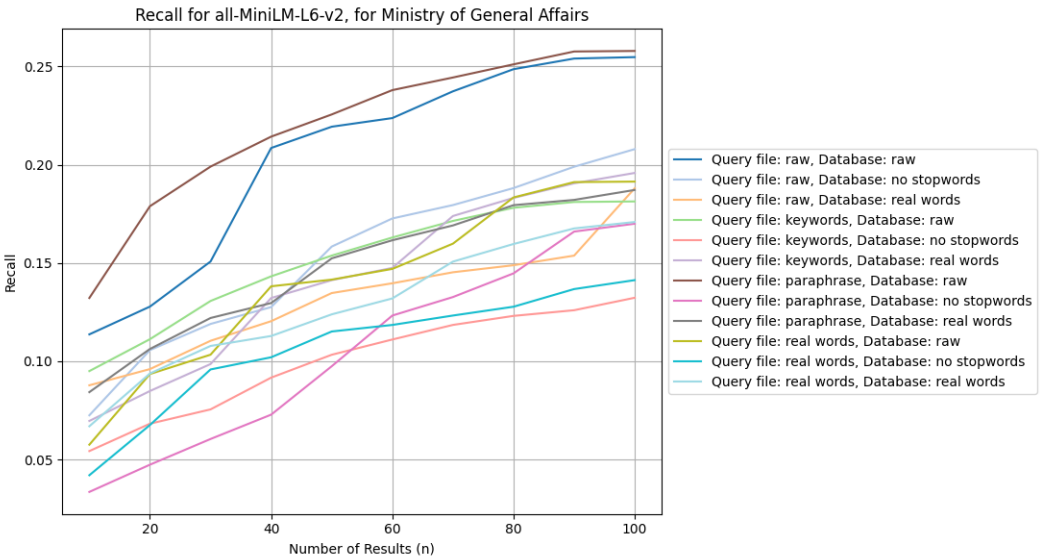

Figure A.18: Dataset: Ministry of General Affairs, Model: MiniLM, Metric: Recall

Table A.19: **Mean Average Precision** at Different Amount of Pages Retrieved Values for **BERTje** on **Ministry of Foreign Affairs**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.0408	0.0289	0.0233	0.0199	0.0173	0.0155	0.0143	0.0133	0.0130	0.0131
	stopwords	0.0035	0.0023	0.0017	0.0013	0.0011	0.0010	0.0009	0.0008	0.0008	0.0007
	real words	0.0085	0.0069	0.0058	0.0050	0.0042	0.0038	0.0035	0.0032	0.0031	0.0030
keywords	raw	0.0192	0.0114	0.0086	0.0069	0.0058	0.0052	0.0049	0.0046	0.0045	0.0050
	stopwords	0.0022	0.0018	0.0015	0.0012	0.0010	0.0009	0.0007	0.0007	0.0006	0.0006
	real words	0.0131	0.0083	0.0065	0.0055	0.0048	0.0044	0.0041	0.0039	0.0037	0.0035
paraphrase	raw	0.0428	0.0283	0.0217	0.0183	0.0161	0.0145	0.0134	0.0127	0.0124	0.0135
	stopwords	0.0028	0.0018	0.0013	0.0011	0.0009	0.0008	0.0007	0.0006	0.0006	0.0005
	real words	0.0216	0.0135	0.0104	0.0085	0.0077	0.0067	0.0060	0.0054	0.0050	0.0046
real words	raw	0.0087	0.0070	0.0053	0.0048	0.0045	0.0041	0.0038	0.0036	0.0035	0.0037
	stopwords	0.0041	0.0030	0.0025	0.0022	0.0019	0.0018	0.0016	0.0015	0.0014	0.0014
	real words	0.0433	0.0299	0.0246	0.0206	0.0176	0.0154	0.0139	0.0126	0.0116	0.0108

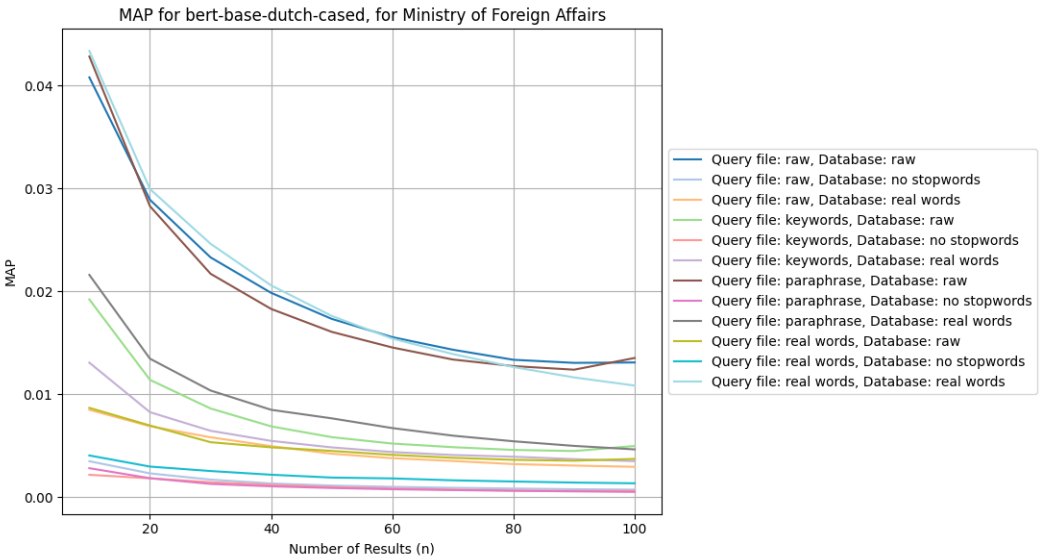

Figure A.19: Dataset: Ministry of Foreign Affairs, Model: BERTje, Metric: MAP

Table A.20: **Precision** at Different Amount of Pages Retrieved Values for **BERTje** on **Ministry of Foreign Affairs**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.0495	0.0395	0.0338	0.0292	0.0257	0.0232	0.0217	0.0201	0.0188	0.0176
	stopwords	0.0061	0.0061	0.0061	0.0057	0.0060	0.0060	0.0061	0.0061	0.0061	0.0061
	real words	0.0173	0.0181	0.0165	0.0151	0.0136	0.0133	0.0130	0.0123	0.0122	0.0120
keywords	raw	0.0316	0.0237	0.0206	0.0194	0.0179	0.0172	0.0173	0.0170	0.0168	0.0155
	stopwords	0.0082	0.0082	0.0080	0.0075	0.0066	0.0060	0.0055	0.0051	0.0054	0.0057
	real words	0.0260	0.0194	0.0187	0.0176	0.0172	0.0172	0.0170	0.0172	0.0170	0.0168
paraphrase	raw	0.0643	0.0510	0.0429	0.0397	0.0372	0.0356	0.0341	0.0324	0.0302	0.0278
	stopwords	0.0066	0.0061	0.0053	0.0055	0.0056	0.0054	0.0055	0.0052	0.0053	0.0049
	real words	0.0316	0.0276	0.0240	0.0232	0.0222	0.0209	0.0200	0.0189	0.0180	0.0174
real words	raw	0.0148	0.0143	0.0133	0.0133	0.0131	0.0133	0.0125	0.0122	0.0120	0.0116
	stopwords	0.0092	0.0084	0.0087	0.0085	0.0092	0.0098	0.0091	0.0091	0.0087	0.0088
	real words	0.0526	0.0411	0.0362	0.0324	0.0297	0.0276	0.0265	0.0252	0.0236	0.0226

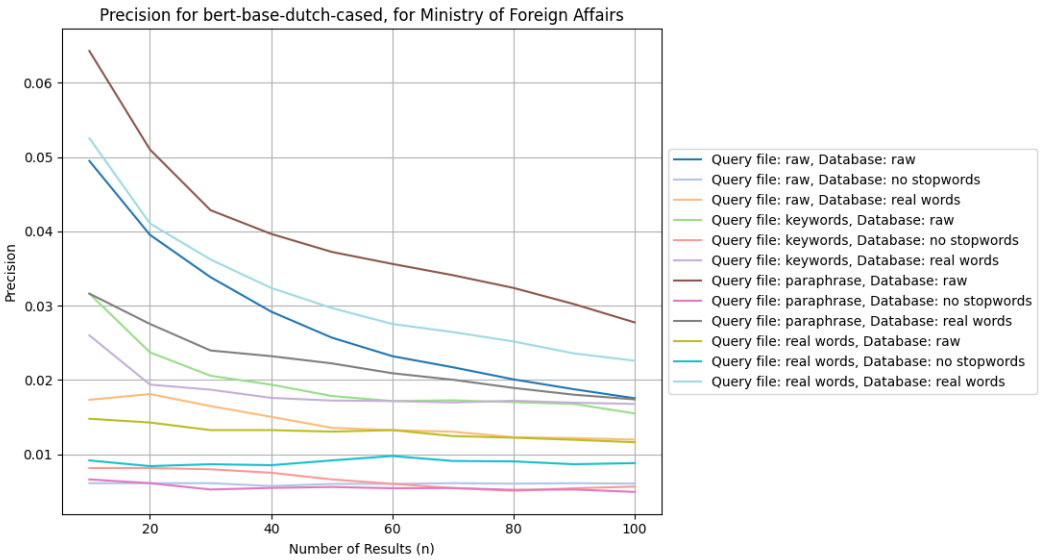

Figure A.20: Dataset: Ministry of Foreign Affairs, Model: BERTje, Metric: Precision

Table A.21: **Recall** at Different Amount of Pages Retrieved Values for **BERTje** on **Ministry of Foreign Affairs**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.0053	0.0080	0.0092	0.0098	0.0104	0.0110	0.0116	0.0122	0.0134	0.0137
	stopwords	0.0005	0.0018	0.0031	0.0032	0.0040	0.0042	0.0050	0.0053	0.0055	0.0058
	real words	0.0036	0.0056	0.0076	0.0083	0.0091	0.0095	0.0100	0.0103	0.0106	0.0109
keywords	raw	0.0027	0.0038	0.0045	0.0062	0.0070	0.0076	0.0086	0.0099	0.0106	0.0109
	stopwords	0.0007	0.0012	0.0018	0.0023	0.0024	0.0026	0.0028	0.0030	0.0034	0.0036
	real words	0.0019	0.0028	0.0046	0.0059	0.0067	0.0078	0.0087	0.0096	0.0107	0.0116
paraphrase	raw	0.0064	0.0091	0.0119	0.0137	0.0157	0.0188	0.0205	0.0224	0.0235	0.0241
	stopwords	0.0006	0.0009	0.0015	0.0021	0.0029	0.0031	0.0038	0.0039	0.0043	0.0043
	real words	0.0040	0.0061	0.0074	0.0089	0.0096	0.0109	0.0136	0.0140	0.0154	0.0164
real words	raw	0.0014	0.0018	0.0027	0.0032	0.0045	0.0057	0.0059	0.0063	0.0067	0.0070
	stopwords	0.0005	0.0008	0.0011	0.0015	0.0030	0.0041	0.0046	0.0053	0.0055	0.0061
	real words	0.0096	0.0127	0.0139	0.0150	0.0160	0.0170	0.0182	0.0189	0.0192	0.0196

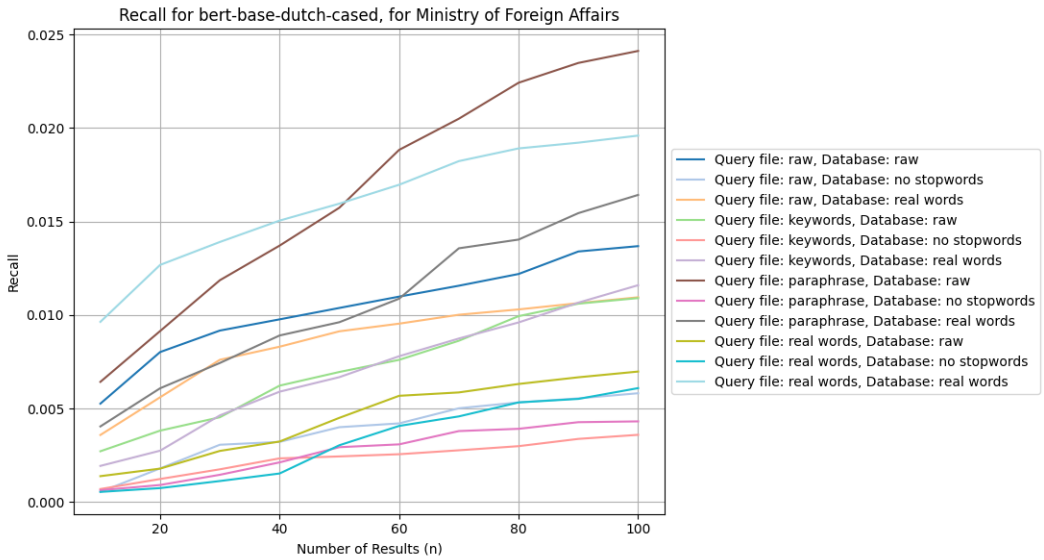

Figure A.21: Dataset: Ministry of Foreign Affairs, Model: BERTje, Metric: Recall

Table A.22: **Mean Average Precision** at Different Amount of Pages Retrieved Values for **BM25** on **Ministry of Foreign Affairs**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.7709	0.7009	0.6486	0.6069	0.5711	0.5392	0.5091	0.4833	0.4580	0.4364
	stopwords	0.7679	0.6953	0.6415	0.5962	0.5582	0.5259	0.4951	0.4690	0.4451	0.4228
	real words	0.6490	0.5844	0.5416	0.5021	0.4682	0.4379	0.4103	0.3858	0.3637	0.3449
keywords	raw	0.2055	0.1613	0.1334	0.1159	0.1030	0.0934	0.0859	0.0800	0.0747	0.0701
	stopwords	0.1673	0.1289	0.1088	0.0965	0.0881	0.0812	0.0755	0.0713	0.0676	0.0642
	real words	0.1317	0.1023	0.0850	0.0735	0.0649	0.0586	0.0535	0.0497	0.0473	0.0450
paraphrase	raw	0.2778	0.2162	0.1828	0.1608	0.1435	0.1307	0.1200	0.1117	0.1041	0.0977
	stopwords	0.2496	0.1962	0.1670	0.1464	0.1322	0.1211	0.1115	0.1039	0.0980	0.0926
	real words	0.1650	0.1298	0.1077	0.0935	0.0832	0.0748	0.0688	0.0637	0.0602	0.0567
real words	raw	0.6224	0.5594	0.5123	0.4737	0.4405	0.4114	0.3852	0.3631	0.3430	0.3257
	stopwords	0.6071	0.5359	0.4870	0.4456	0.4123	0.3815	0.3559	0.3336	0.3140	0.2974
	real words	0.6592	0.5986	0.5531	0.5144	0.4786	0.4475	0.4194	0.3951	0.3740	0.3549

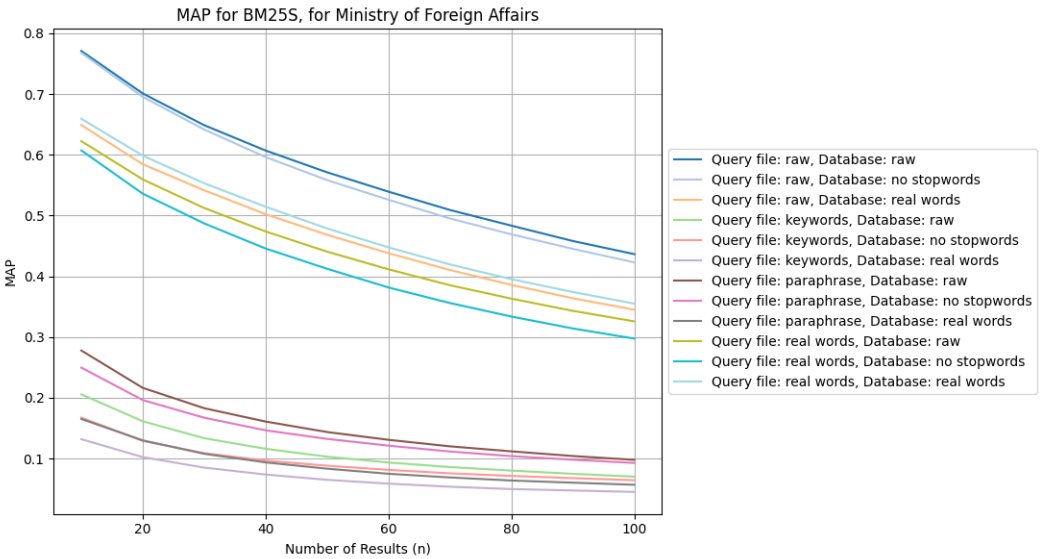

Figure A.22: Dataset: Ministry of Foreign Affairs, Model: BM25, Metric: MAP

Table A.23: **Precision** at Different Amount of Pages Retrieved Values for **BM25** on **Ministry of Foreign Affairs**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.7944	0.7309	0.6818	0.6426	0.6096	0.5790	0.5499	0.5247	0.4992	0.4782
	stopwords	0.7954	0.7283	0.6781	0.6348	0.5981	0.5663	0.5359	0.5112	0.4879	0.4658
	real words	0.6847	0.6268	0.5855	0.5503	0.5199	0.4921	0.4655	0.4410	0.4184	0.4003
keywords	raw	0.2418	0.2059	0.1781	0.1611	0.1486	0.1395	0.1326	0.1267	0.1210	0.1155
	stopwords	0.2102	0.1719	0.1531	0.1407	0.1332	0.1258	0.1184	0.1133	0.1089	0.1048
	real words	0.1673	0.1441	0.1276	0.1152	0.1059	0.0992	0.0929	0.0888	0.0875	0.0851
paraphrase	raw	0.3301	0.2747	0.2439	0.2228	0.2040	0.1913	0.1792	0.1710	0.1618	0.1542
	stopwords	0.2944	0.2469	0.2175	0.1969	0.1833	0.1727	0.1640	0.1562	0.1502	0.1440
	real words	0.2122	0.1855	0.1617	0.1478	0.1368	0.1261	0.1193	0.1129	0.1091	0.1051
real words	raw	0.6526	0.5987	0.5571	0.5240	0.4941	0.4662	0.4409	0.4195	0.3995	0.3833
	stopwords	0.6469	0.5827	0.5396	0.5005	0.4704	0.4403	0.4157	0.3937	0.3738	0.3572
	real words	0.6888	0.6349	0.5929	0.5598	0.5267	0.4973	0.4712	0.4474	0.4266	0.4080

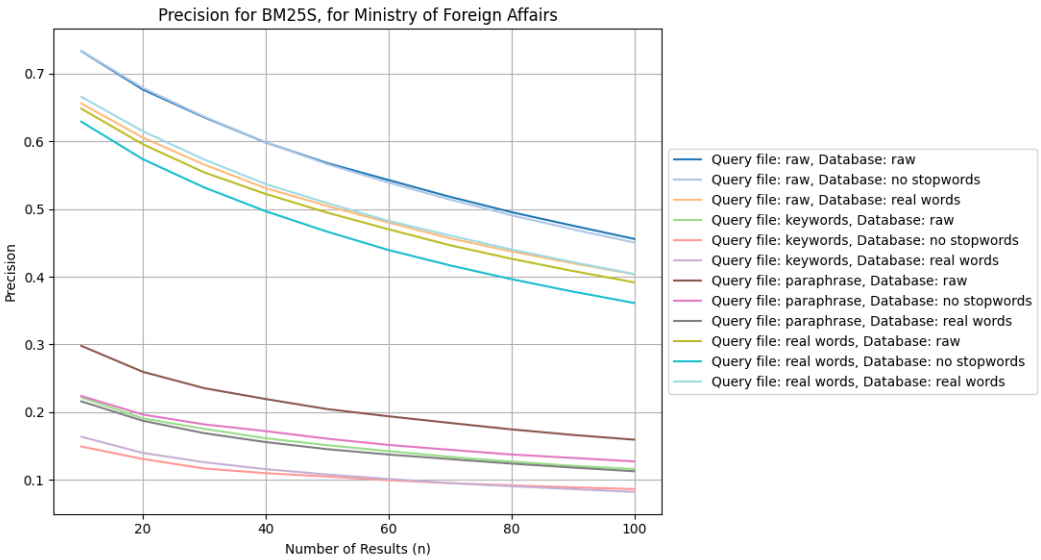

Figure A.23: Dataset: Ministry of Foreign Affairs, Model: BM25, Metric: Precision

Table A.24: **Recall** at Different Amount of Pages Retrieved Values for **BM25** on **Ministry of Foreign Affairs**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.2050	0.2959	0.3577	0.4071	0.4485	0.4834	0.5109	0.5313	0.5487	0.5630
	stopwords	0.2037	0.2948	0.3593	0.4048	0.4423	0.4690	0.4903	0.5123	0.5293	0.5436
	real words	0.1363	0.2075	0.2533	0.2932	0.3288	0.3632	0.3868	0.4040	0.4178	0.4339
keywords	raw	0.0530	0.0783	0.0910	0.1015	0.1099	0.1172	0.1282	0.1337	0.1395	0.1462
	stopwords	0.0489	0.0653	0.0764	0.0869	0.0963	0.1021	0.1062	0.1114	0.1167	0.1229
	real words	0.0335	0.0486	0.0588	0.0659	0.0792	0.0851	0.0889	0.0930	0.0982	0.1025
paraphrase	raw	0.0858	0.1129	0.1333	0.1466	0.1609	0.1764	0.1866	0.1969	0.2019	0.2076
	stopwords	0.0760	0.1036	0.1199	0.1356	0.1486	0.1602	0.1704	0.1771	0.1842	0.1893
	real words	0.0536	0.0756	0.0887	0.1034	0.1130	0.1177	0.1253	0.1312	0.1390	0.1435
real words	raw	0.1209	0.1819	0.2319	0.2695	0.3012	0.3249	0.3455	0.3659	0.3804	0.3993
	stopwords	0.1252	0.1862	0.2298	0.2635	0.2942	0.3162	0.3333	0.3472	0.3618	0.3759
	real words	0.1334	0.2050	0.2536	0.2978	0.3341	0.3635	0.3910	0.4097	0.4251	0.4420

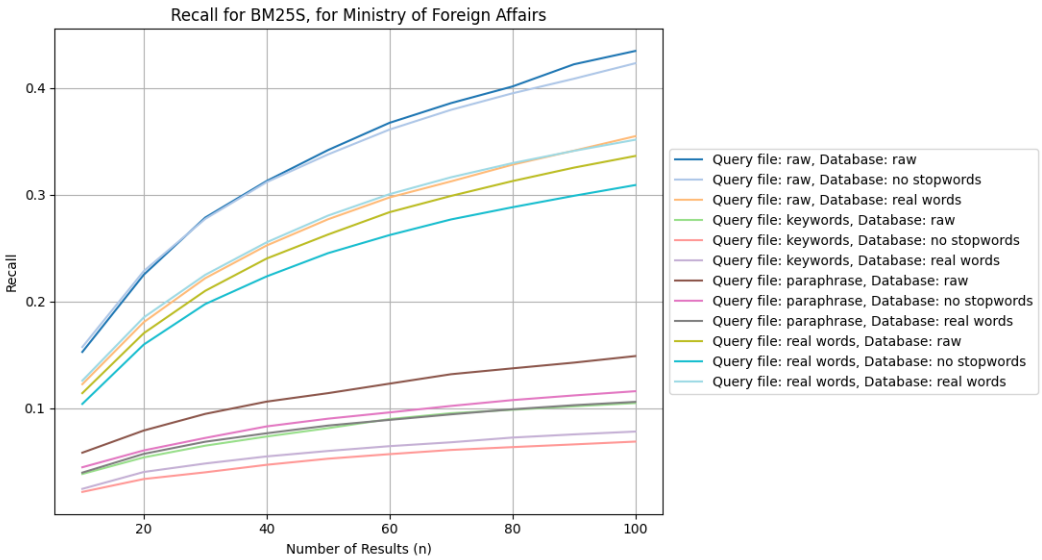

Figure A.24: Dataset: Ministry of Foreign Affairs, Model: BM25, Metric: Recall

Table A.25: **Mean Average Precision** at Different Amount of Pages Retrieved Values for **MiniLM** on **Ministry of Foreign Affairs**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.1111	0.0843	0.0711	0.0608	0.0539	0.0493	0.0461	0.0434	0.0417	0.0425
	stopwords	0.0571	0.0408	0.0346	0.0296	0.0263	0.0235	0.0213	0.0196	0.0183	0.0173
	real words	0.0855	0.0616	0.0517	0.0454	0.0413	0.0376	0.0346	0.0324	0.0307	0.0292
keywords	raw	0.0922	0.0727	0.0626	0.0558	0.0507	0.0467	0.0439	0.0422	0.0415	0.0430
	stopwords	0.0486	0.0386	0.0328	0.0291	0.0261	0.0241	0.0223	0.0208	0.0197	0.0185
	real words	0.0620	0.0482	0.0409	0.0361	0.0331	0.0306	0.0292	0.0277	0.0265	0.0257
paraphrase	raw	0.1117	0.0846	0.0710	0.0638	0.0580	0.0537	0.0500	0.0472	0.0456	0.0469
	stopwords	0.0668	0.0527	0.0440	0.0392	0.0356	0.0325	0.0299	0.0281	0.0265	0.0251
	real words	0.0687	0.0520	0.0452	0.0404	0.0372	0.0346	0.0328	0.0310	0.0298	0.0287
real words	raw	0.0548	0.0382	0.0308	0.0268	0.0233	0.0212	0.0195	0.0185	0.0182	0.0195
	stopwords	0.0274	0.0209	0.0181	0.0155	0.0140	0.0123	0.0112	0.0103	0.0095	0.0089
	real words	0.0823	0.0574	0.0444	0.0372	0.0325	0.0288	0.0261	0.0237	0.0220	0.0206

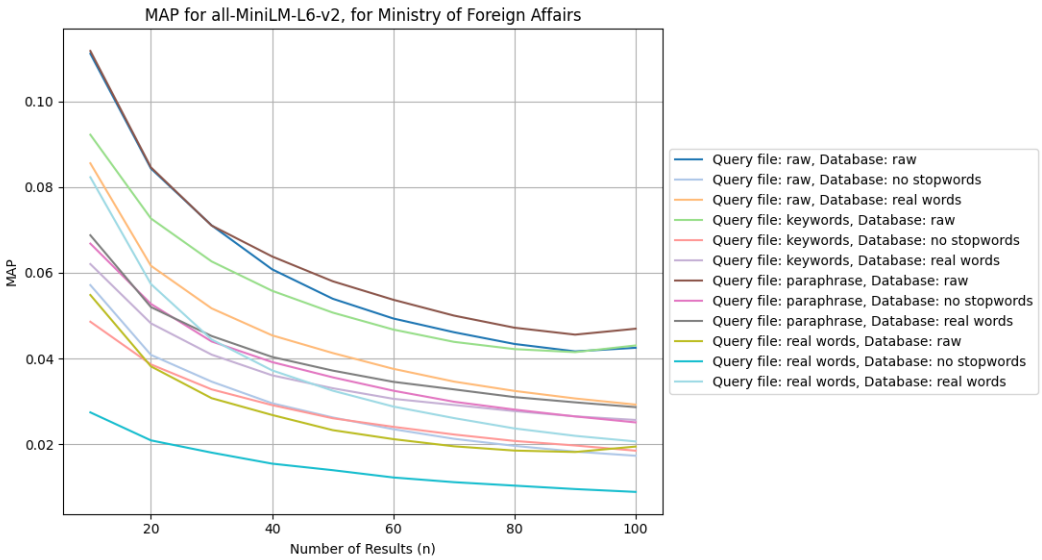

Figure A.25: Dataset: Ministry of Foreign Affairs, Model: MiniLM, Metric: MAP

Table A.26: **Precision** at Different Amount of Pages Retrieved Values for **MiniLM** on **Ministry of Foreign Affairs**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.1429	0.1196	0.1078	0.0980	0.0901	0.0853	0.0800	0.0760	0.0715	0.0663
	stopwords	0.0827	0.0673	0.0604	0.0551	0.0509	0.0478	0.0452	0.0437	0.0422	0.0411
	real words	0.1128	0.0908	0.0840	0.0784	0.0744	0.0700	0.0659	0.0633	0.0616	0.0599
keywords	raw	0.1219	0.1107	0.1036	0.0954	0.0897	0.0851	0.0821	0.0797	0.0766	0.0702
	stopwords	0.0709	0.0620	0.0578	0.0541	0.0506	0.0487	0.0463	0.0441	0.0426	0.0411
	real words	0.0913	0.0798	0.0748	0.0698	0.0680	0.0648	0.0637	0.0620	0.0599	0.0587
paraphrase	raw	0.1561	0.1321	0.1170	0.1094	0.1024	0.0974	0.0927	0.0878	0.0830	0.0762
	stopwords	0.0939	0.0816	0.0707	0.0665	0.0636	0.0599	0.0561	0.0541	0.0527	0.0508
	real words	0.1036	0.0893	0.0835	0.0787	0.0744	0.0707	0.0686	0.0659	0.0642	0.0623
real words	raw	0.0776	0.0630	0.0561	0.0531	0.0484	0.0469	0.0447	0.0429	0.0410	0.0377
	stopwords	0.0449	0.0390	0.0357	0.0327	0.0309	0.0286	0.0273	0.0265	0.0253	0.0244
	real words	0.1102	0.0849	0.0713	0.0638	0.0588	0.0541	0.0509	0.0480	0.0459	0.0444

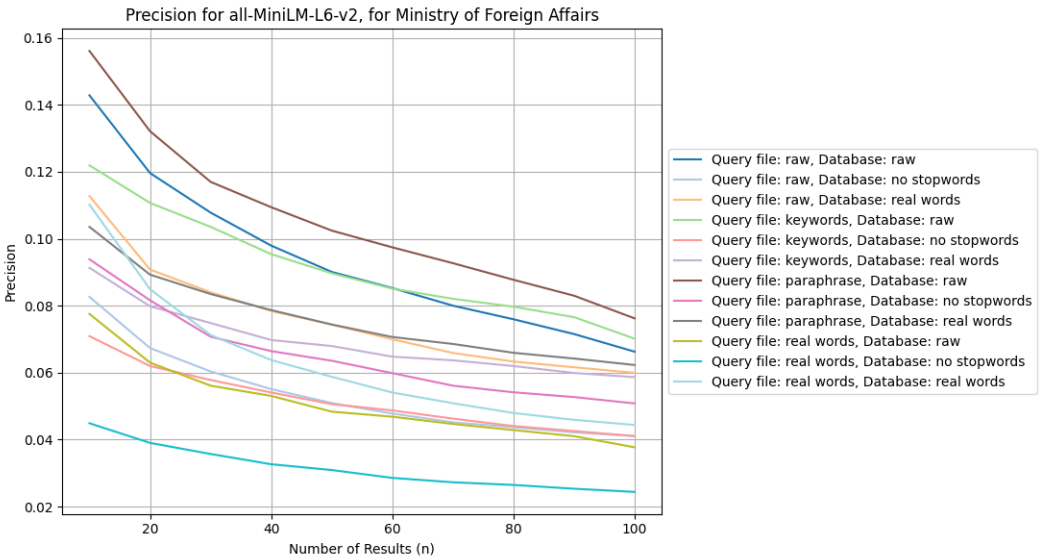

Figure A.26: Dataset: Ministry of Foreign Affairs, Model: MiniLM, Metric: Precision

Table A.27: **Recall** at Different Amount of Pages Retrieved Values for **MiniLM** on **Ministry of Foreign Affairs**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.0217	0.0309	0.0389	0.0460	0.0520	0.0566	0.0599	0.0649	0.0677	0.0687
	stopwords	0.0101	0.0149	0.0189	0.0215	0.0241	0.0266	0.0294	0.0317	0.0337	0.0351
	real words	0.0102	0.0163	0.0237	0.0300	0.0346	0.0383	0.0414	0.0447	0.0484	0.0513
keywords	raw	0.0164	0.0255	0.0340	0.0450	0.0514	0.0554	0.0601	0.0638	0.0672	0.0692
	stopwords	0.0069	0.0111	0.0143	0.0164	0.0191	0.0215	0.0234	0.0246	0.0260	0.0279
	real words	0.0089	0.0160	0.0207	0.0249	0.0290	0.0322	0.0363	0.0391	0.0419	0.0439
paraphrase	raw	0.0275	0.0365	0.0439	0.0523	0.0602	0.0656	0.0695	0.0733	0.0767	0.0775
	stopwords	0.0107	0.0159	0.0185	0.0216	0.0251	0.0273	0.0288	0.0311	0.0330	0.0346
	real words	0.0137	0.0206	0.0292	0.0336	0.0378	0.0438	0.0472	0.0512	0.0548	0.0570
real words	raw	0.0138	0.0210	0.0246	0.0279	0.0298	0.0319	0.0342	0.0364	0.0382	0.0386
	stopwords	0.0051	0.0079	0.0100	0.0117	0.0138	0.0148	0.0160	0.0180	0.0193	0.0208
	real words	0.0151	0.0214	0.0244	0.0282	0.0310	0.0329	0.0351	0.0370	0.0388	0.0405

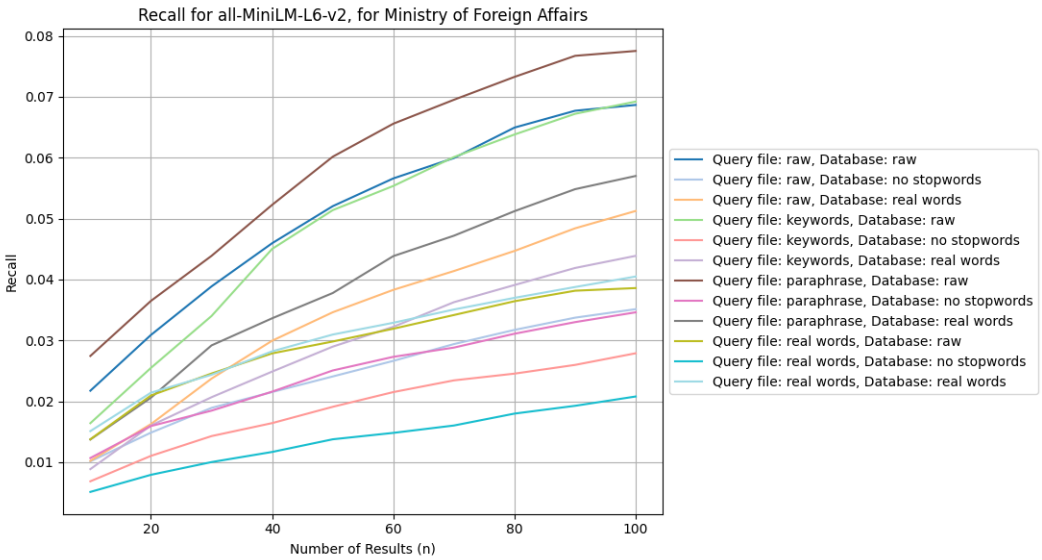

Figure A.27: Dataset: Ministry of Foreign Affairs, Model: MiniLM, Metric: Recall

Table A.28: **Mean Average Precision** at Different Amount of Pages Retrieved Values for **BERTje** on **Ministry of Finance**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.0164	0.0103	0.0075	0.0060	0.0051	0.0045	0.0041	0.0039	0.0037	0.0039
	stopwords	0.0020	0.0013	0.0010	0.0008	0.0007	0.0006	0.0006	0.0006	0.0005	0.0005
	real words	0.0073	0.0043	0.0031	0.0026	0.0023	0.0021	0.0019	0.0018	0.0016	0.0016
keywords	raw	0.0063	0.0041	0.0032	0.0027	0.0023	0.0021	0.0020	0.0019	0.0018	0.0020
	stopwords	0.0035	0.0022	0.0016	0.0013	0.0011	0.0010	0.0008	0.0008	0.0007	0.0007
	real words	0.0064	0.0040	0.0031	0.0025	0.0022	0.0019	0.0017	0.0015	0.0014	0.0013
paraphrase	raw	0.0357	0.0247	0.0202	0.0175	0.0154	0.0137	0.0125	0.0116	0.0112	0.0120
	stopwords	0.0027	0.0017	0.0012	0.0010	0.0009	0.0008	0.0007	0.0006	0.0006	0.0006
	real words	0.0082	0.0061	0.0051	0.0042	0.0037	0.0034	0.0032	0.0030	0.0027	0.0026
real words	raw	0.0047	0.0042	0.0042	0.0042	0.0042	0.0042	0.0042	0.0040	0.0041	0.0040
	stopwords	0.0014	0.0010	0.0008	0.0007	0.0006	0.0006	0.0006	0.0005	0.0006	0.0006
	real words	0.0186	0.0116	0.0093	0.0080	0.0071	0.0064	0.0059	0.0054	0.0050	0.0046

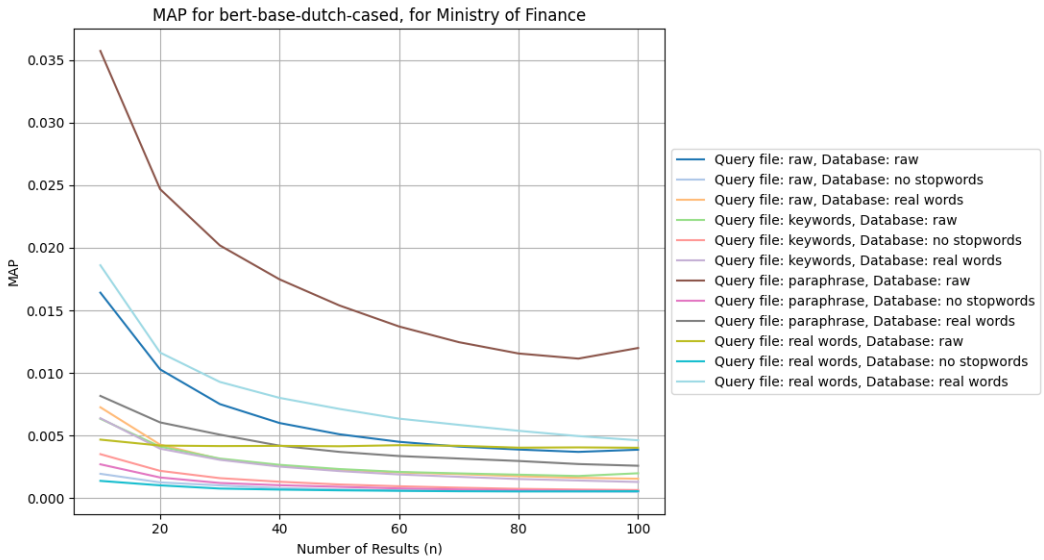

Figure A.28: Dataset: Ministry of Finance, Model: BERTje, Metric: MAP

Table A.29: **Precision** at Different Amount of Pages Retrieved Values for **BERTje** on **Ministry of Finance**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.0218	0.0174	0.0151	0.0133	0.0126	0.0120	0.0110	0.0106	0.0104	0.0097
	stopwords	0.0036	0.0039	0.0041	0.0038	0.0039	0.0038	0.0038	0.0040	0.0040	0.0040
	real words	0.0106	0.0088	0.0078	0.0078	0.0079	0.0077	0.0077	0.0074	0.0071	0.0073
keywords	raw	0.0146	0.0129	0.0116	0.0114	0.0111	0.0108	0.0106	0.0103	0.0099	0.0092
	stopwords	0.0065	0.0055	0.0051	0.0053	0.0049	0.0048	0.0044	0.0041	0.0040	0.0040
	real words	0.0126	0.0118	0.0108	0.0103	0.0096	0.0091	0.0086	0.0082	0.0080	0.0077
paraphrase	raw	0.0561	0.0453	0.0415	0.0384	0.0359	0.0331	0.0314	0.0297	0.0284	0.0261
	stopwords	0.0065	0.0054	0.0044	0.0045	0.0045	0.0042	0.0040	0.0040	0.0041	0.0041
	real words	0.0184	0.0177	0.0164	0.0154	0.0144	0.0141	0.0142	0.0139	0.0133	0.0131
real words	raw	0.0070	0.0064	0.0067	0.0068	0.0066	0.0068	0.0066	0.0066	0.0066	0.0066
	stopwords	0.0045	0.0046	0.0041	0.0041	0.0040	0.0041	0.0041	0.0041	0.0041	0.0043
	real words	0.0249	0.0178	0.0159	0.0149	0.0139	0.0133	0.0132	0.0126	0.0120	0.0117

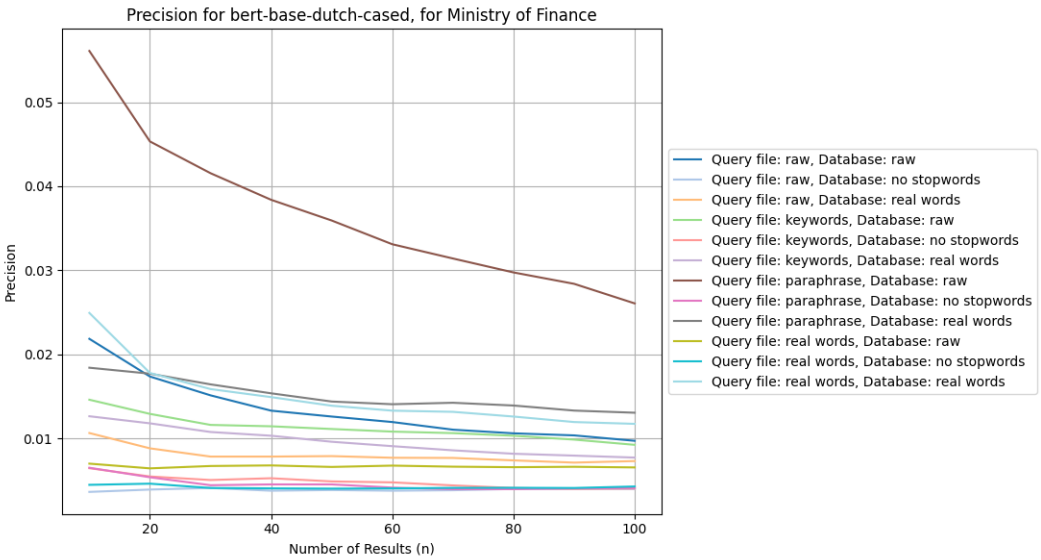

Figure A.29: Dataset: Ministry of Finance, Model: BERTje, Metric: Precision

Table A.30: **Recall** at Different Amount of Pages Retrieved Values for **BERTje** on **Ministry of Finance**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.0056	0.0065	0.0069	0.0081	0.0088	0.0098	0.0101	0.0105	0.0114	0.0117
	stopwords	0.0002	0.0008	0.0010	0.0012	0.0020	0.0022	0.0027	0.0029	0.0031	0.0035
	real words	0.0014	0.0019	0.0027	0.0037	0.0043	0.0049	0.0052	0.0057	0.0060	0.0067
keywords	raw	0.0013	0.0022	0.0033	0.0041	0.0047	0.0055	0.0090	0.0095	0.0100	0.0102
	stopwords	0.0004	0.0011	0.0017	0.0019	0.0020	0.0023	0.0024	0.0025	0.0026	0.0027
	real words	0.0011	0.0019	0.0025	0.0035	0.0041	0.0046	0.0050	0.0054	0.0059	0.0063
paraphrase	raw	0.0065	0.0104	0.0130	0.0144	0.0164	0.0174	0.0189	0.0206	0.0223	0.0227
	stopwords	0.0003	0.0005	0.0005	0.0007	0.0009	0.0010	0.0011	0.0013	0.0015	0.0018
	real words	0.0012	0.0029	0.0045	0.0053	0.0063	0.0070	0.0086	0.0094	0.0103	0.0114
real words	raw	0.0003	0.0006	0.0008	0.0011	0.0012	0.0015	0.0016	0.0018	0.0020	0.0022
	stopwords	0.0002	0.0008	0.0010	0.0011	0.0014	0.0016	0.0019	0.0022	0.0023	0.0025
	real words	0.0066	0.0076	0.0089	0.0098	0.0113	0.0126	0.0140	0.0145	0.0151	0.0161

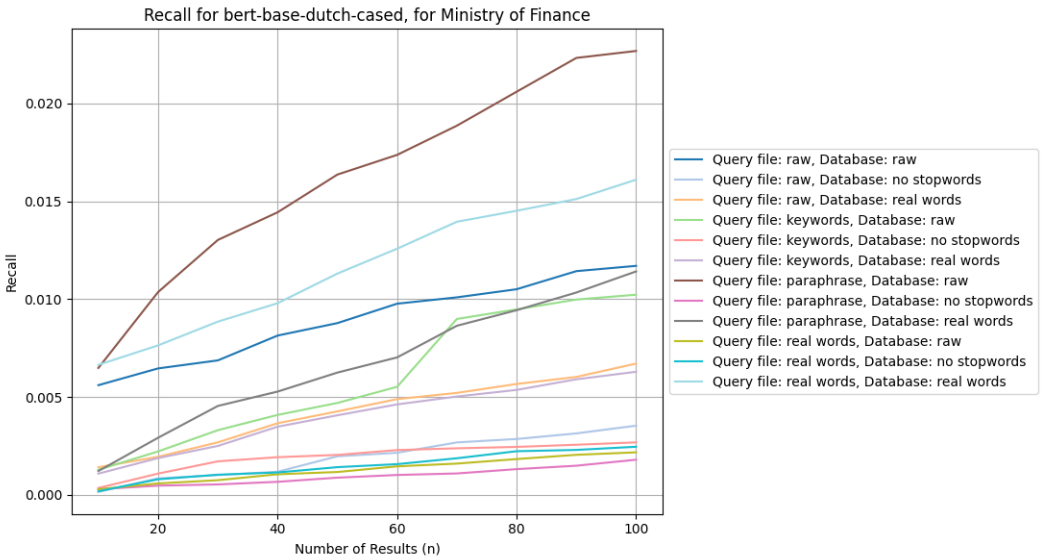

Figure A.30: Dataset: Ministry of Finance, Model: BERTje, Metric: Recall

Table A.31: **Mean Average Precision** at Different Amount of Pages Retrieved Values for **BM25** on **Ministry of Finance**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.7676	0.6819	0.6250	0.5745	0.5338	0.4993	0.4707	0.4456	0.4232	0.4030
	stopwords	0.7645	0.6764	0.6142	0.5617	0.5202	0.4844	0.4545	0.4294	0.4073	0.3872
	real words	0.6899	0.6070	0.5457	0.4960	0.4563	0.4236	0.3958	0.3722	0.3518	0.3330
keywords	raw	0.1366	0.1121	0.0981	0.0885	0.0804	0.0737	0.0681	0.0633	0.0593	0.0558
	stopwords	0.0984	0.0802	0.0695	0.0617	0.0557	0.0515	0.0481	0.0449	0.0423	0.0400
	real words	0.0871	0.0685	0.0589	0.0521	0.0464	0.0421	0.0386	0.0356	0.0331	0.0311
paraphrase	raw	0.2490	0.1947	0.1653	0.1460	0.1301	0.1172	0.1068	0.0982	0.0910	0.0848
	stopwords	0.1907	0.1487	0.1260	0.1103	0.0977	0.0885	0.0808	0.0743	0.0693	0.0649
	real words	0.1428	0.1043	0.0856	0.0734	0.0643	0.0573	0.0522	0.0478	0.0442	0.0409
real words	raw	0.6975	0.6126	0.5512	0.5020	0.4611	0.4279	0.3996	0.3748	0.3530	0.3338
	stopwords	0.6751	0.5800	0.5173	0.4671	0.4273	0.3945	0.3675	0.3436	0.3226	0.3036
	real words	0.7137	0.6281	0.5673	0.5176	0.4778	0.4430	0.4141	0.3889	0.3663	0.3463

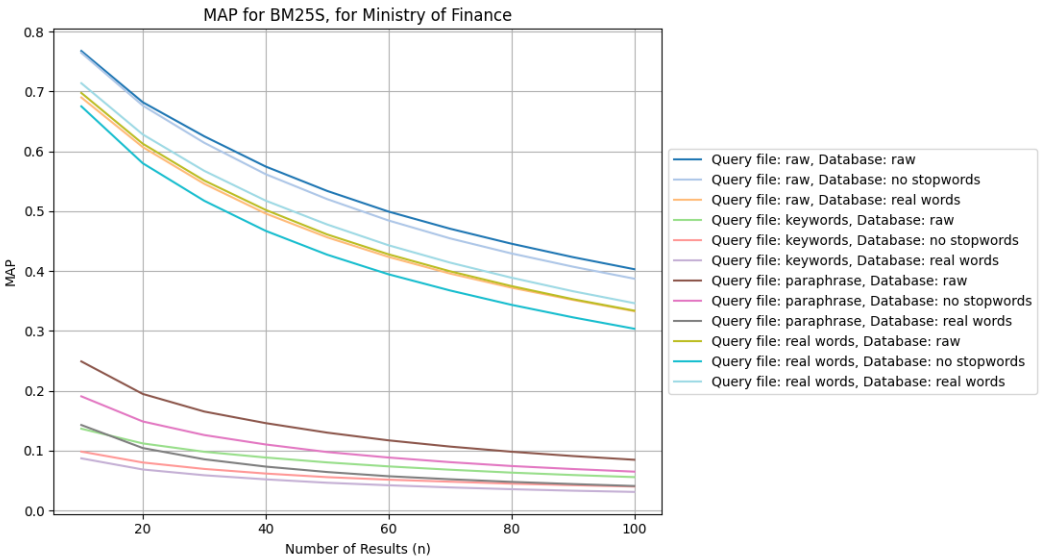

Figure A.31: Dataset: Ministry of Finance, Model: BM25, Metric: MAP

Table A.32: **Precision** at Different Amount of Pages Retrieved Values for **BM25** on **Ministry of Finance**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.8003	0.7197	0.6681	0.6183	0.5781	0.5433	0.5149	0.4897	0.4671	0.4472
	stopwords	0.7961	0.7150	0.6563	0.6062	0.5661	0.5302	0.5004	0.4751	0.4529	0.4327
	real words	0.7294	0.6583	0.6010	0.5520	0.5125	0.4799	0.4518	0.4274	0.4066	0.3871
keywords	raw	0.1728	0.1518	0.1385	0.1301	0.1218	0.1136	0.1074	0.1019	0.0972	0.0931
	stopwords	0.1301	0.1139	0.1029	0.0952	0.0898	0.0859	0.0825	0.0785	0.0753	0.0720
	real words	0.1146	0.1000	0.0915	0.0841	0.0771	0.0720	0.0674	0.0639	0.0607	0.0583
paraphrase	raw	0.2966	0.2480	0.2209	0.2028	0.1857	0.1706	0.1585	0.1481	0.1397	0.1321
	stopwords	0.2365	0.2020	0.1802	0.1647	0.1504	0.1397	0.1303	0.1225	0.1163	0.1105
	real words	0.1816	0.1490	0.1302	0.1163	0.1069	0.0980	0.0926	0.0867	0.0820	0.0774
real words	raw	0.7359	0.6618	0.6039	0.5568	0.5165	0.4833	0.4543	0.4289	0.4063	0.3867
	stopwords	0.7190	0.6304	0.5723	0.5242	0.4852	0.4524	0.4246	0.3995	0.3779	0.3577
	real words	0.7499	0.6734	0.6179	0.5702	0.5312	0.4954	0.4662	0.4406	0.4177	0.3971

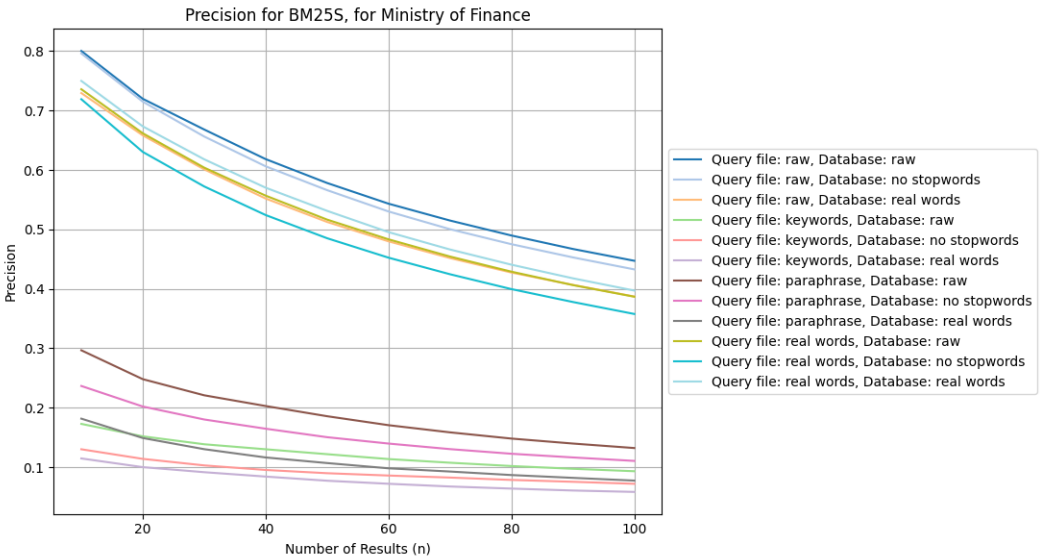

Figure A.32: Dataset: Ministry of Finance, Model: BM25, Metric: Precision

Table A.33: **Recall** at Different Amount of Pages Retrieved Values for **BM25** on **Ministry of Finance**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.2540	0.3516	0.4259	0.4733	0.5121	0.5397	0.5636	0.5839	0.6034	0.6202
	stopwords	0.2485	0.3465	0.4150	0.4617	0.4972	0.5234	0.5471	0.5664	0.5833	0.5998
	real words	0.2033	0.2975	0.3591	0.3984	0.4294	0.4587	0.4812	0.4988	0.5155	0.5292
keywords	raw	0.0468	0.0646	0.0756	0.0918	0.1033	0.1096	0.1150	0.1212	0.1261	0.1316
	stopwords	0.0315	0.0451	0.0533	0.0611	0.0699	0.0772	0.0821	0.0857	0.0894	0.0919
	real words	0.0319	0.0430	0.0506	0.0579	0.0630	0.0687	0.0721	0.0772	0.0800	0.0828
paraphrase	raw	0.0723	0.1035	0.1265	0.1455	0.1589	0.1676	0.1751	0.1805	0.1867	0.1919
	stopwords	0.0555	0.0827	0.1005	0.1129	0.1248	0.1327	0.1383	0.1459	0.1508	0.1559
	real words	0.0482	0.0657	0.0772	0.0870	0.0947	0.0996	0.1063	0.1099	0.1137	0.1166
real words	raw	0.2024	0.2932	0.3526	0.3995	0.4339	0.4583	0.4791	0.4981	0.5141	0.5275
	stopwords	0.1978	0.2747	0.3322	0.3738	0.4020	0.4288	0.4504	0.4662	0.4805	0.4927
	real words	0.2113	0.3029	0.3670	0.4094	0.4451	0.4719	0.4940	0.5127	0.5282	0.5418

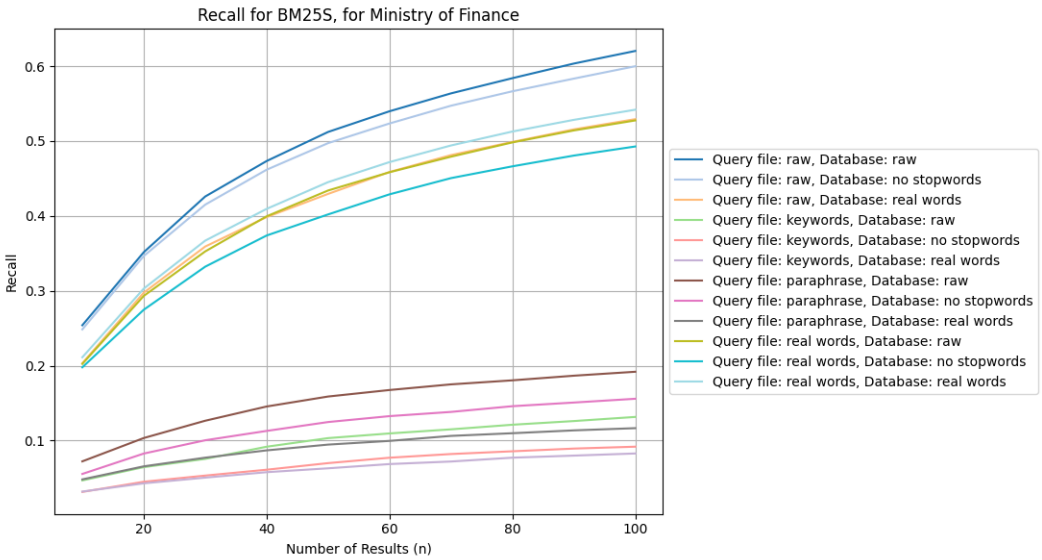

Figure A.33: Dataset: Ministry of Finance, Model: BM25, Metric: Recall

Table A.34: **Mean Average Precision** at Different Amount of Pages Retrieved Values for **MiniLM** on **Ministry of Finance**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.0436	0.0318	0.0257	0.0218	0.0189	0.0165	0.0150	0.0138	0.0131	0.0132
	stopwords	0.0176	0.0127	0.0102	0.0089	0.0081	0.0073	0.0066	0.0061	0.0056	0.0053
	real words	0.0289	0.0185	0.0147	0.0123	0.0107	0.0095	0.0083	0.0075	0.0068	0.0063
keywords	raw	0.0370	0.0267	0.0218	0.0191	0.0174	0.0159	0.0147	0.0140	0.0137	0.0141
	stopwords	0.0141	0.0106	0.0095	0.0086	0.0079	0.0074	0.0070	0.0068	0.0063	0.0059
	real words	0.0184	0.0139	0.0114	0.0099	0.0087	0.0076	0.0069	0.0063	0.0058	0.0054
paraphrase	raw	0.0580	0.0430	0.0362	0.0313	0.0275	0.0245	0.0225	0.0210	0.0203	0.0207
	stopwords	0.0157	0.0120	0.0104	0.0097	0.0087	0.0081	0.0076	0.0072	0.0068	0.0064
	real words	0.0204	0.0158	0.0136	0.0120	0.0110	0.0101	0.0094	0.0086	0.0081	0.0077
real words	raw	0.0211	0.0154	0.0128	0.0112	0.0103	0.0091	0.0082	0.0076	0.0072	0.0073
	stopwords	0.0100	0.0064	0.0050	0.0042	0.0038	0.0033	0.0031	0.0028	0.0026	0.0024
	real words	0.0323	0.0213	0.0170	0.0140	0.0121	0.0106	0.0094	0.0085	0.0077	0.0071

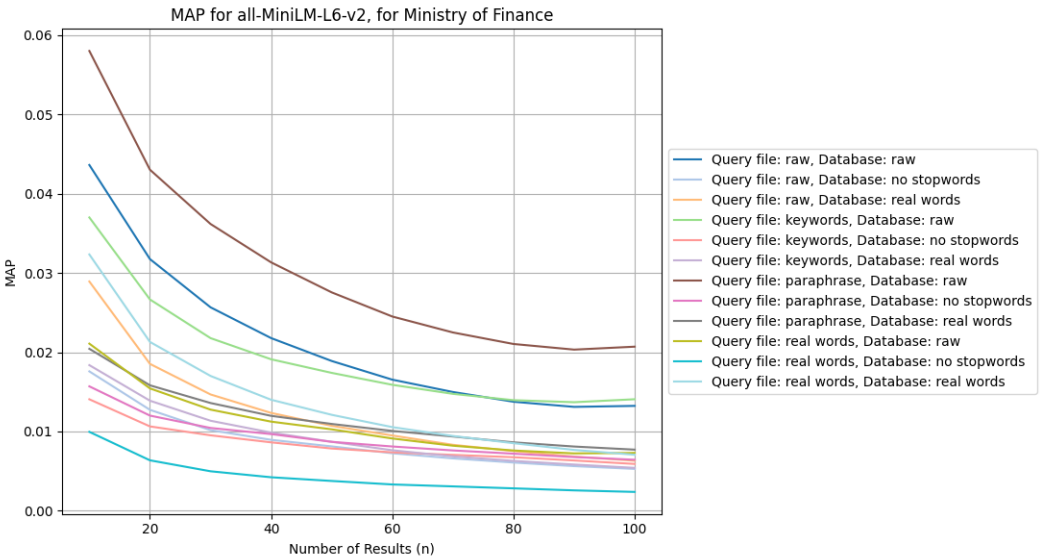

Figure A.34: Dataset: Ministry of Finance, Model: MiniLM, Metric: MAP

Table A.35: **Precision** at Different Amount of Pages Retrieved Values for **MiniLM** on **Ministry of Finance**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.0613	0.0503	0.0443	0.0394	0.0367	0.0337	0.0317	0.0293	0.0272	0.0255
	stopwords	0.0261	0.0214	0.0183	0.0173	0.0167	0.0155	0.0144	0.0138	0.0131	0.0128
	real words	0.0429	0.0332	0.0294	0.0270	0.0253	0.0242	0.0224	0.0213	0.0202	0.0192
keywords	raw	0.0517	0.0430	0.0390	0.0364	0.0350	0.0327	0.0313	0.0296	0.0277	0.0256
	stopwords	0.0256	0.0219	0.0225	0.0213	0.0196	0.0190	0.0186	0.0180	0.0173	0.0165
	real words	0.0309	0.0277	0.0243	0.0230	0.0216	0.0201	0.0188	0.0180	0.0172	0.0166
paraphrase	raw	0.0813	0.0677	0.0597	0.0547	0.0497	0.0458	0.0430	0.0408	0.0386	0.0360
	stopwords	0.0278	0.0248	0.0243	0.0240	0.0223	0.0216	0.0208	0.0203	0.0195	0.0189
	real words	0.0348	0.0317	0.0279	0.0256	0.0246	0.0239	0.0226	0.0216	0.0206	0.0199
real words	raw	0.0308	0.0261	0.0227	0.0217	0.0210	0.0195	0.0185	0.0174	0.0162	0.0153
	stopwords	0.0151	0.0120	0.0107	0.0099	0.0096	0.0087	0.0089	0.0084	0.0080	0.0078
	real words	0.0445	0.0331	0.0287	0.0249	0.0229	0.0210	0.0198	0.0188	0.0176	0.0167

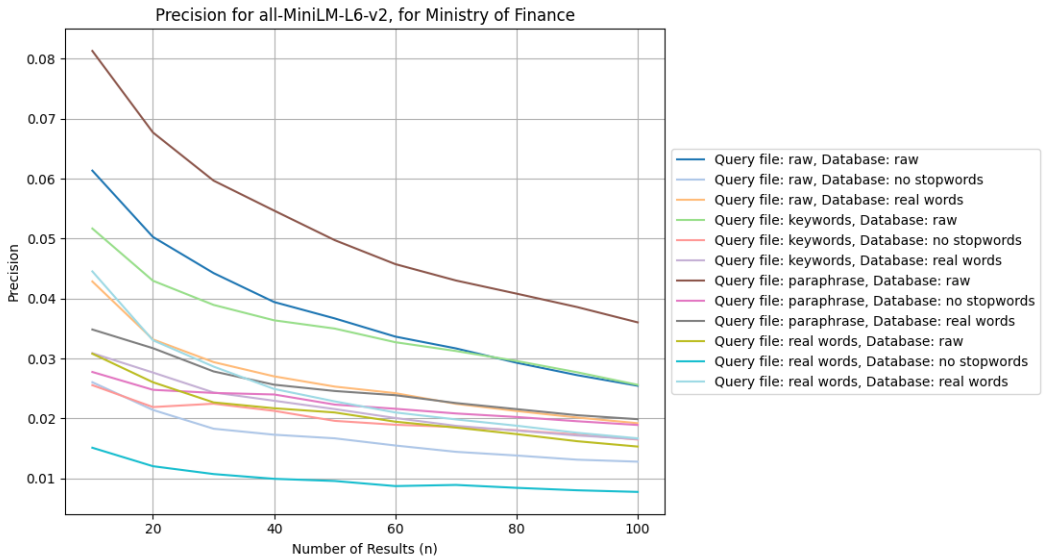

Figure A.35: Dataset: Ministry of Finance, Model: MiniLM, Metric: Precision

Table A.36: **Recall** at Different Amount of Pages Retrieved Values for **MiniLM** on **Ministry of Finance**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.0116	0.0156	0.0192	0.0212	0.0239	0.0257	0.0274	0.0287	0.0297	0.0307
	stopwords	0.0061	0.0078	0.0087	0.0097	0.0110	0.0120	0.0134	0.0144	0.0149	0.0158
	real words	0.0082	0.0109	0.0131	0.0152	0.0172	0.0190	0.0228	0.0238	0.0247	0.0256
keywords	raw	0.0103	0.0145	0.0177	0.0202	0.0227	0.0247	0.0263	0.0283	0.0294	0.0300
	stopwords	0.0046	0.0060	0.0084	0.0099	0.0105	0.0118	0.0135	0.0150	0.0158	0.0163
	real words	0.0061	0.0084	0.0100	0.0114	0.0127	0.0139	0.0147	0.0156	0.0168	0.0177
paraphrase	raw	0.0141	0.0199	0.0238	0.0279	0.0294	0.0319	0.0336	0.0355	0.0368	0.0380
	stopwords	0.0062	0.0073	0.0085	0.0101	0.0113	0.0123	0.0136	0.0144	0.0152	0.0162
	real words	0.0032	0.0058	0.0099	0.0145	0.0166	0.0179	0.0185	0.0194	0.0201	0.0208
real words	raw	0.0070	0.0089	0.0104	0.0123	0.0134	0.0144	0.0154	0.0162	0.0166	0.0170
	stopwords	0.0022	0.0030	0.0036	0.0046	0.0053	0.0055	0.0060	0.0091	0.0100	0.0104
	real words	0.0091	0.0116	0.0144	0.0156	0.0170	0.0182	0.0191	0.0199	0.0206	0.0212

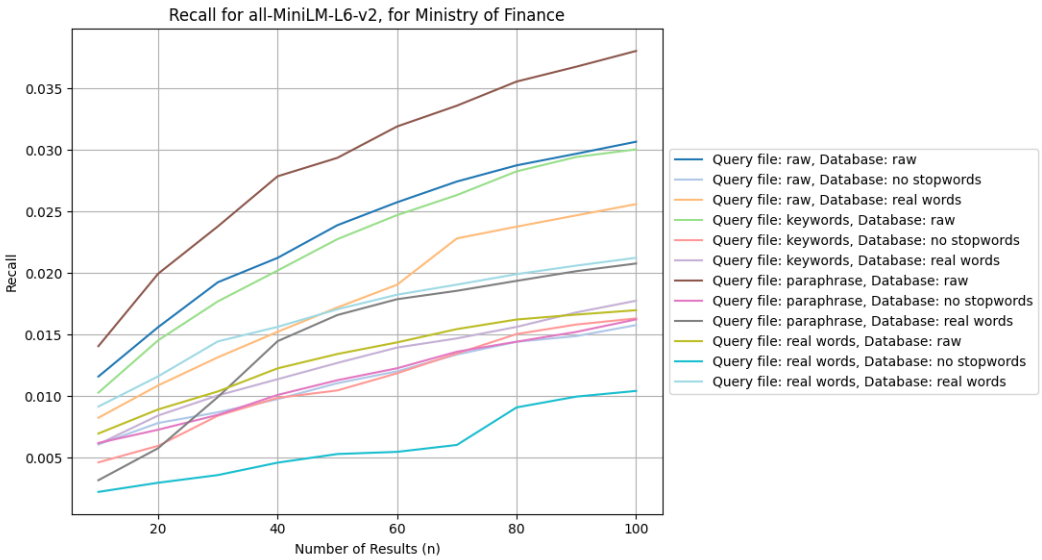

Figure A.36: Dataset: Ministry of Finance, Model: MiniLM, Metric: Recall

Table A.37: **Mean Average Precision** at Different Amount of Pages Retrieved Values for **BERTje** on **Ministry of Justice and Safety**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.0272	0.0171	0.0134	0.0115	0.0102	0.0093	0.0084	0.0078	0.0074	0.0073
	stopwords	0.0015	0.0019	0.0019	0.0018	0.0014	0.0012	0.0011	0.0009	0.0008	0.0007
	real words	0.0065	0.0039	0.0029	0.0023	0.0019	0.0017	0.0015	0.0014	0.0013	0.0012
keywords	raw	0.0037	0.0030	0.0024	0.0020	0.0018	0.0017	0.0016	0.0015	0.0015	0.0018
	stopwords	0.0012	0.0012	0.0008	0.0006	0.0005	0.0005	0.0006	0.0006	0.0007	0.0007
	real words	0.0033	0.0023	0.0019	0.0016	0.0014	0.0012	0.0011	0.0010	0.0009	0.0009
paraphrase	raw	0.0292	0.0217	0.0181	0.0160	0.0144	0.0129	0.0120	0.0113	0.0111	0.0117
	stopwords	0.0021	0.0012	0.0008	0.0006	0.0005	0.0004	0.0004	0.0004	0.0003	0.0003
	real words	0.0141	0.0093	0.0075	0.0063	0.0057	0.0052	0.0048	0.0044	0.0041	0.0039
real words	raw	0.0066	0.0054	0.0048	0.0044	0.0041	0.0038	0.0037	0.0035	0.0035	0.0037
	stopwords	0.0021	0.0021	0.0019	0.0019	0.0020	0.0019	0.0018	0.0017	0.0015	0.0014
	real words	0.0189	0.0123	0.0089	0.0072	0.0059	0.0052	0.0046	0.0042	0.0039	0.0036

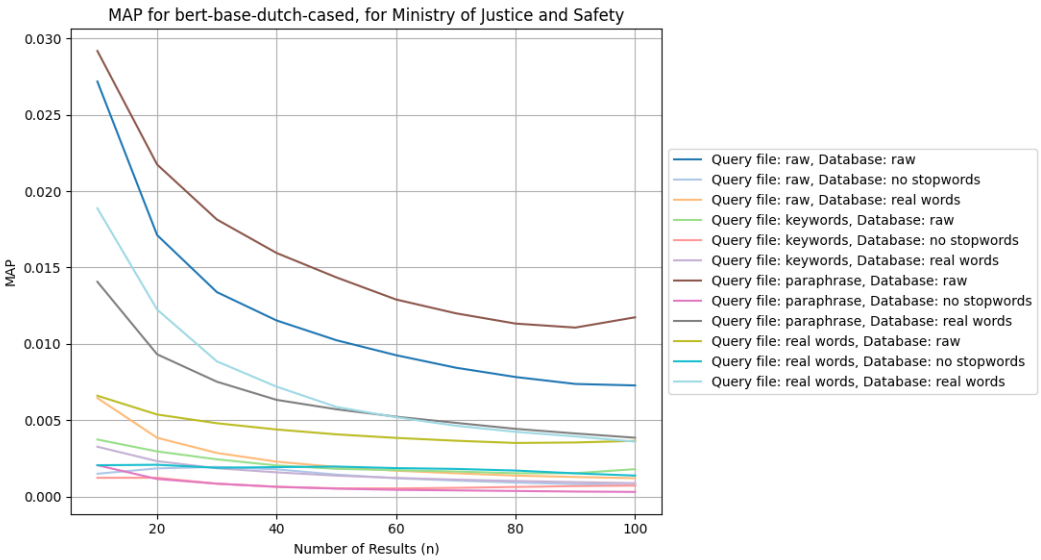

Figure A.37: Dataset: Ministry of Justice and Safety, Model: BERTje, Metric: MAP

Table A.38: **Precision** at Different Amount of Pages Retrieved Values for **BERTje** on **Ministry of Ministry of Justice and Safety**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.0346	0.0242	0.0200	0.0182	0.0168	0.0156	0.0144	0.0135	0.0128	0.0119
	stopwords	0.0032	0.0033	0.0031	0.0028	0.0024	0.0022	0.0022	0.0020	0.0018	0.0018
	real words	0.0090	0.0076	0.0069	0.0061	0.0060	0.0059	0.0058	0.0055	0.0056	0.0055
keywords	raw	0.0084	0.0086	0.0081	0.0078	0.0075	0.0078	0.0075	0.0074	0.0069	0.0066
	stopwords	0.0035	0.0036	0.0028	0.0024	0.0023	0.0026	0.0027	0.0029	0.0030	0.0032
	real words	0.0086	0.0083	0.0080	0.0084	0.0080	0.0079	0.0075	0.0073	0.0071	0.0069
paraphrase	raw	0.0425	0.0359	0.0330	0.0310	0.0292	0.0273	0.0261	0.0249	0.0236	0.0219
	stopwords	0.0047	0.0038	0.0034	0.0027	0.0023	0.0023	0.0023	0.0022	0.0021	0.0020
	real words	0.0251	0.0212	0.0205	0.0190	0.0180	0.0176	0.0167	0.0157	0.0151	0.0147
real words	raw	0.0095	0.0079	0.0070	0.0065	0.0065	0.0062	0.0061	0.0060	0.0060	0.0058
	stopwords	0.0030	0.0031	0.0031	0.0033	0.0037	0.0035	0.0034	0.0032	0.0030	0.0028
	real words	0.0224	0.0172	0.0135	0.0121	0.0109	0.0106	0.0102	0.0098	0.0094	0.0089

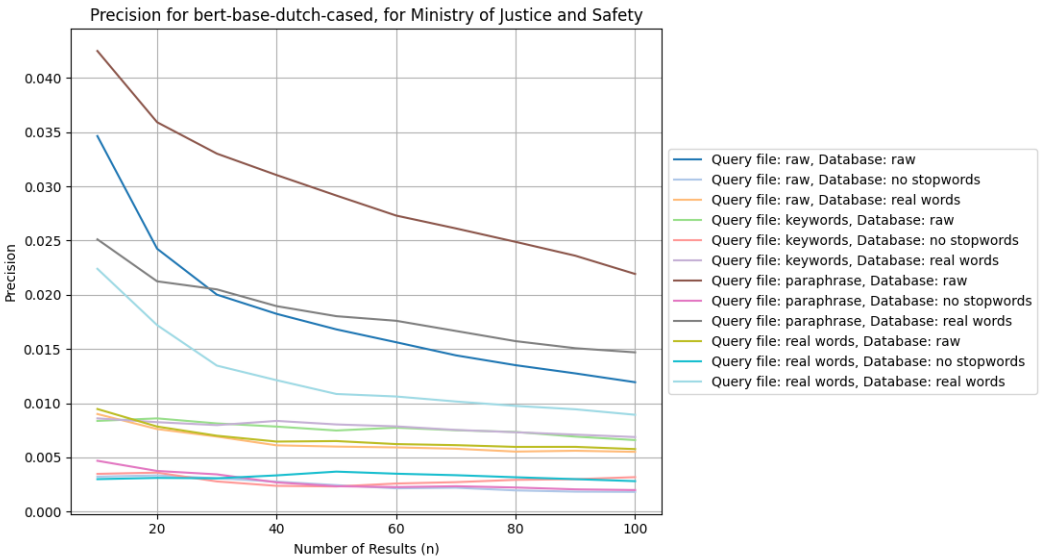

Figure A.38: Dataset: Ministry of Justice and Safety, Model: BERTje, Metric: Precision

Table A.39: **Recall** at Different Amount of Pages Retrieved Values for **BERTje** on **Ministry of Ministry of Justice and Safety**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.0112	0.0119	0.0123	0.0132	0.0140	0.0144	0.0146	0.0153	0.0156	0.0157
	stopwords	0.0001	0.0002	0.0003	0.0003	0.0004	0.0005	0.0006	0.0006	0.0006	0.0007
	real words	0.0020	0.0024	0.0028	0.0042	0.0046	0.0048	0.0056	0.0057	0.0061	0.0063
keywords	raw	0.0032	0.0038	0.0040	0.0043	0.0046	0.0049	0.0050	0.0055	0.0055	0.0058
	stopwords	0.0005	0.0012	0.0013	0.0013	0.0014	0.0018	0.0019	0.0020	0.0021	0.0022
	real words	0.0012	0.0020	0.0027	0.0032	0.0036	0.0044	0.0048	0.0053	0.0056	0.0057
paraphrase	raw	0.0040	0.0059	0.0074	0.0086	0.0106	0.0112	0.0123	0.0130	0.0137	0.0140
	stopwords	0.0001	0.0002	0.0004	0.0004	0.0005	0.0006	0.0006	0.0006	0.0007	0.0007
	real words	0.0021	0.0031	0.0045	0.0054	0.0059	0.0069	0.0072	0.0076	0.0082	0.0088
real words	raw	0.0030	0.0033	0.0036	0.0038	0.0040	0.0042	0.0045	0.0046	0.0048	0.0049
	stopwords	0.0001	0.0001	0.0002	0.0003	0.0004	0.0005	0.0005	0.0006	0.0006	0.0007
	real words	0.0068	0.0077	0.0079	0.0086	0.0089	0.0091	0.0095	0.0096	0.0097	0.0099

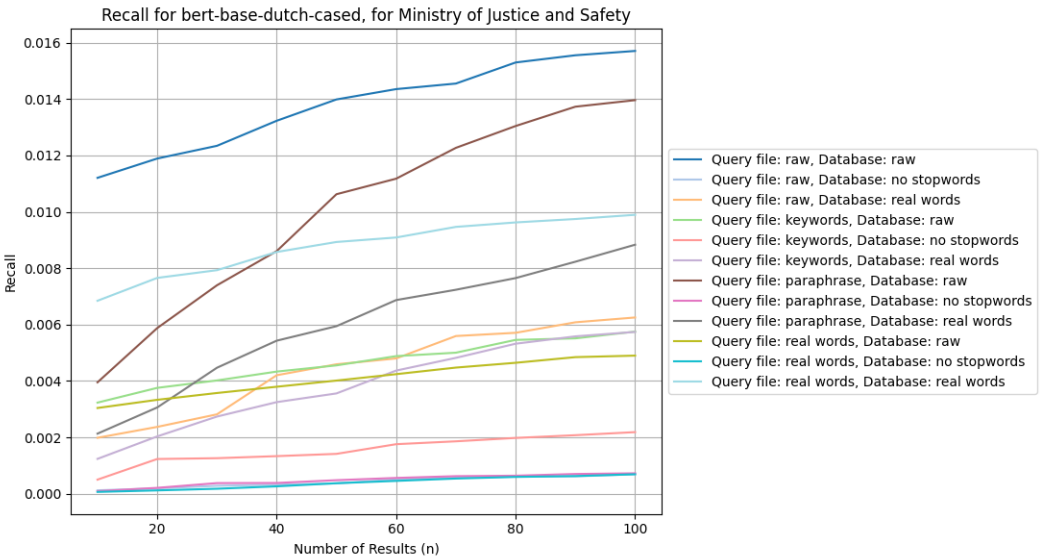

Figure A.39: Dataset: Ministry of Justice and Safety, Model: BERTje, Metric: Recall

Table A.40: **Mean Average Precision** at Different Amount of Pages Retrieved Values for **BM25** on **Ministry of Ministry of Justice and Safety**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.7036	0.6412	0.5969	0.5590	0.5276	0.5019	0.4769	0.4543	0.4341	0.4148
	stopwords	0.7022	0.6421	0.5966	0.5579	0.5252	0.4975	0.4720	0.4488	0.4281	0.4085
	real words	0.6228	0.5643	0.5206	0.4846	0.4570	0.4319	0.4089	0.3897	0.3720	0.3553
keywords	raw	0.1843	0.1512	0.1337	0.1200	0.1102	0.1019	0.0945	0.0884	0.0831	0.0790
	stopwords	0.1208	0.0984	0.0846	0.0765	0.0710	0.0662	0.0624	0.0594	0.0568	0.0545
	real words	0.1344	0.1075	0.0933	0.0833	0.0761	0.0701	0.0649	0.0606	0.0567	0.0534
paraphrase	raw	0.2528	0.2069	0.1808	0.1651	0.1512	0.1405	0.1314	0.1232	0.1163	0.1102
	stopwords	0.1865	0.1506	0.1341	0.1228	0.1126	0.1044	0.0978	0.0920	0.0872	0.0827
	real words	0.1743	0.1397	0.1213	0.1091	0.0994	0.0922	0.0864	0.0809	0.0761	0.0717
real words	raw	0.6155	0.5576	0.5118	0.4783	0.4500	0.4253	0.4020	0.3820	0.3637	0.3469
	stopwords	0.5935	0.5306	0.4842	0.4477	0.4166	0.3898	0.3670	0.3468	0.3287	0.3122
	real words	0.6315	0.5754	0.5307	0.4936	0.4646	0.4381	0.4162	0.3956	0.3767	0.3593

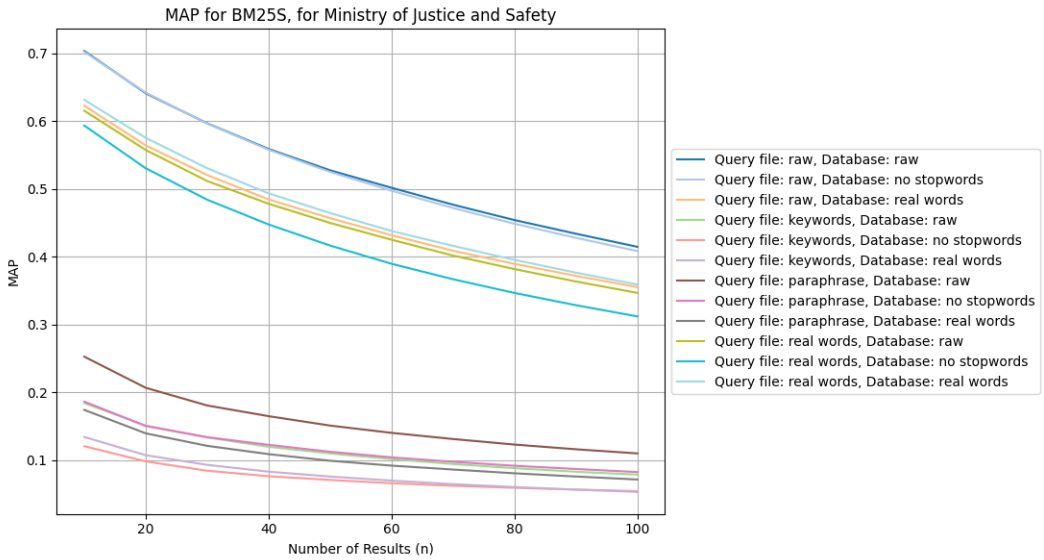

Figure A.40: Dataset: Ministry of Justice and Safety, Model: BM25, Metric: MAP

Table A.41: **Precision** at Different Amount of Pages Retrieved Values for **BM25** on **Ministry of Ministry of Justice and Safety**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.7330	0.6763	0.6353	0.5982	0.5678	0.5427	0.5179	0.4953	0.4753	0.4558
	stopwords	0.7333	0.6788	0.6363	0.5984	0.5662	0.5390	0.5139	0.4906	0.4699	0.4505
	real words	0.6557	0.6054	0.5655	0.5307	0.5045	0.4800	0.4566	0.4373	0.4199	0.4034
keywords	raw	0.2219	0.1909	0.1754	0.1615	0.1513	0.1425	0.1341	0.1271	0.1209	0.1161
	stopwords	0.1493	0.1309	0.1169	0.1099	0.1050	0.0996	0.0953	0.0920	0.0890	0.0865
	real words	0.1637	0.1399	0.1261	0.1159	0.1080	0.1014	0.0953	0.0907	0.0865	0.0823
paraphrase	raw	0.2979	0.2596	0.2354	0.2194	0.2046	0.1939	0.1841	0.1746	0.1664	0.1594
	stopwords	0.2239	0.1965	0.1820	0.1721	0.1609	0.1516	0.1445	0.1375	0.1325	0.1273
	real words	0.2160	0.1873	0.1690	0.1559	0.1454	0.1375	0.1307	0.1241	0.1183	0.1128
real words	raw	0.6480	0.5956	0.5540	0.5222	0.4946	0.4700	0.4464	0.4265	0.4084	0.3916
	stopwords	0.6289	0.5736	0.5318	0.4967	0.4665	0.4394	0.4167	0.3963	0.3780	0.3611
	real words	0.6656	0.6148	0.5730	0.5368	0.5087	0.4824	0.4608	0.4401	0.4215	0.4039

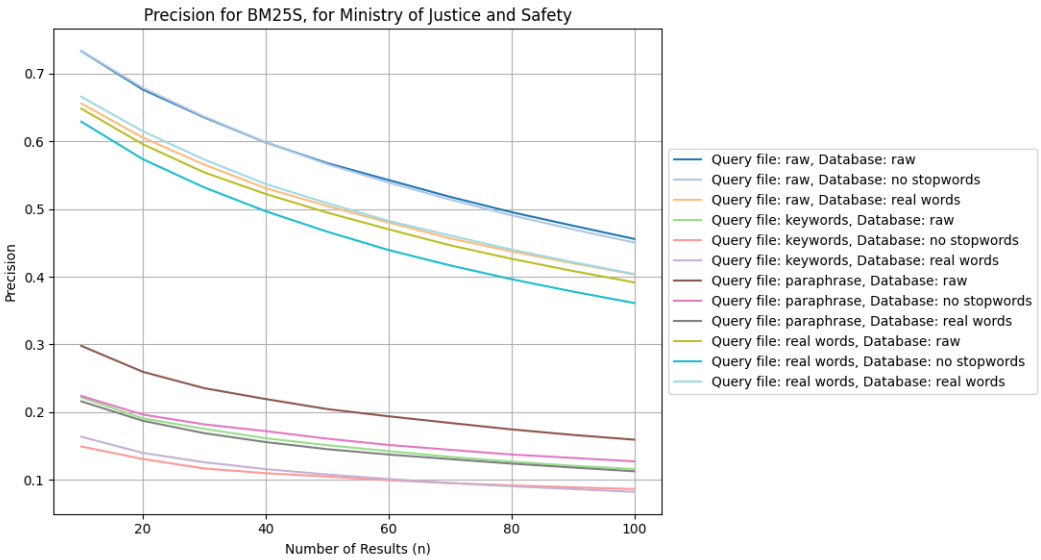

Figure A.41: Dataset: Ministry of Justice and Safety, Model: BM25, Metric: Precision

Table A.42: **Recall** at Different Amount of Pages Retrieved Values for **BM25** on **Ministry of Ministry of Justice and Safety**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.1528	0.2252	0.2788	0.3130	0.3420	0.3676	0.3860	0.4016	0.4224	0.4349
	stopwords	0.1574	0.2286	0.2775	0.3120	0.3380	0.3612	0.3797	0.3952	0.4088	0.4234
	real words	0.1226	0.1807	0.2219	0.2526	0.2772	0.2976	0.3127	0.3283	0.3414	0.3550
keywords	raw	0.0384	0.0540	0.0650	0.0736	0.0814	0.0899	0.0954	0.0986	0.1019	0.1049
	stopwords	0.0217	0.0337	0.0400	0.0471	0.0528	0.0571	0.0610	0.0636	0.0662	0.0688
	real words	0.0245	0.0403	0.0483	0.0549	0.0600	0.0645	0.0682	0.0726	0.0756	0.0782
paraphrase	raw	0.0584	0.0792	0.0948	0.1063	0.1142	0.1231	0.1319	0.1374	0.1428	0.1490
	stopwords	0.0447	0.0605	0.0722	0.0831	0.0903	0.0961	0.1022	0.1077	0.1121	0.1160
	real words	0.0397	0.0573	0.0687	0.0766	0.0838	0.0891	0.0943	0.0991	0.1029	0.1060
real words	raw	0.1142	0.1704	0.2101	0.2404	0.2629	0.2838	0.2990	0.3129	0.3255	0.3365
	stopwords	0.1041	0.1598	0.1977	0.2236	0.2454	0.2623	0.2770	0.2884	0.2991	0.3092
	real words	0.1259	0.1852	0.2251	0.2556	0.2806	0.3007	0.3165	0.3298	0.3413	0.3517

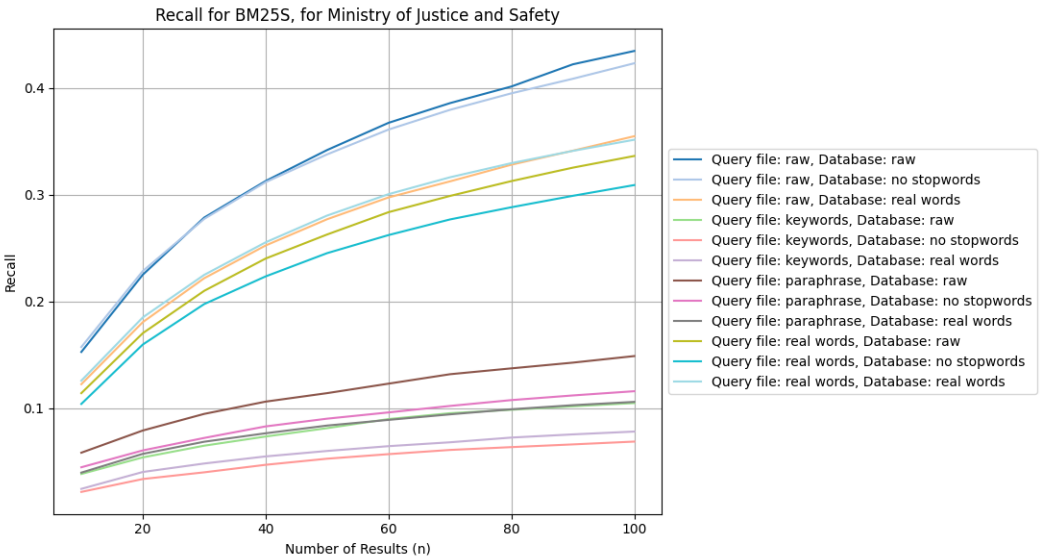

Figure A.42: Dataset: Ministry of Justice and Safety, Model: BM25, Metric: Recall

Table A.43: **Mean Average Precision** at Different Amount of Pages Retrieved Values for **MiniLM** on **Ministry of Ministry of Justice and Safety**

Query File	Database	MAP@10	MAP@20	MAP@30	MAP@40	MAP@50	MAP@60	MAP@70	MAP@80	MAP@90	MAP@100
raw	raw	0.0491	0.0340	0.0279	0.0246	0.0220	0.0202	0.0186	0.0173	0.0165	0.0164
	stopwords	0.0155	0.0110	0.0089	0.0076	0.0068	0.0061	0.0055	0.0051	0.0048	0.0045
	real words	0.0275	0.0197	0.0161	0.0140	0.0122	0.0108	0.0098	0.0090	0.0083	0.0078
keywords	raw	0.0479	0.0390	0.0341	0.0309	0.0284	0.0264	0.0249	0.0239	0.0231	0.0233
	stopwords	0.0172	0.0141	0.0118	0.0108	0.0098	0.0094	0.0088	0.0085	0.0084	0.0082
	real words	0.0279	0.0217	0.0178	0.0158	0.0142	0.0129	0.0122	0.0114	0.0107	0.0101
paraphrase	raw	0.0445	0.0358	0.0307	0.0277	0.0254	0.0236	0.0224	0.0214	0.0207	0.0208
	stopwords	0.0211	0.0160	0.0138	0.0121	0.0112	0.0104	0.0099	0.0092	0.0087	0.0082
	real words	0.0285	0.0198	0.0167	0.0145	0.0130	0.0121	0.0111	0.0105	0.0099	0.0094
real words	raw	0.0177	0.0123	0.0103	0.0090	0.0080	0.0073	0.0069	0.0065	0.0063	0.0064
	stopwords	0.0080	0.0063	0.0056	0.0046	0.0041	0.0037	0.0034	0.0031	0.0030	0.0028
	real words	0.0273	0.0164	0.0124	0.0104	0.0089	0.0078	0.0069	0.0062	0.0057	0.0053

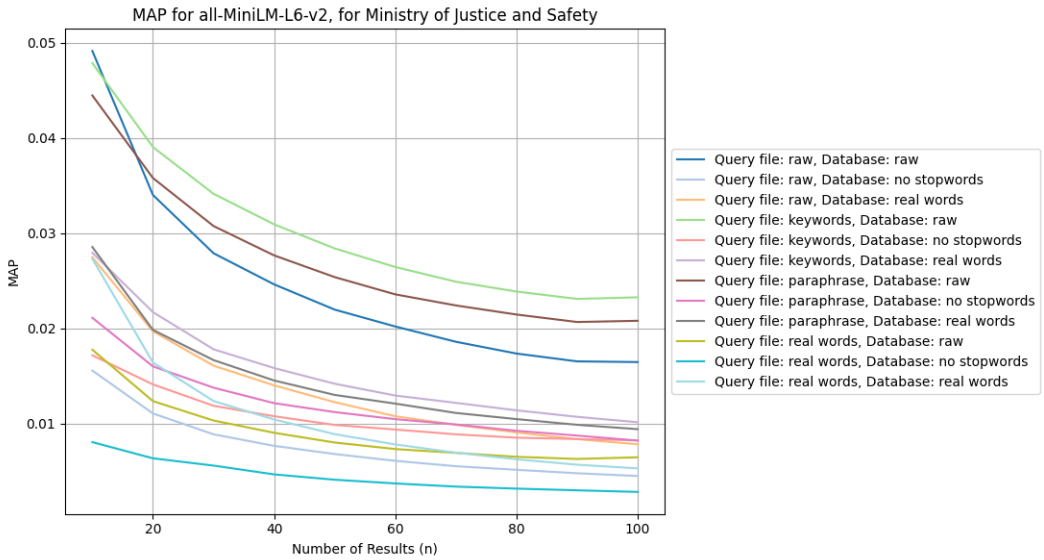

Figure A.43: Dataset: Ministry of Justice and Safety, Model: MiniLM, Metric: MAP

Table A.44: **Precision** at Different Amount of Pages Retrieved Values for **MiniLM** on **Ministry of Ministry of Justice and Safety**

Query File	Database	Precision@10	Precision@20	Precision@30	Precision@40	Precision@50	Precision@60	Precision@70	Precision@80	Precision@90	Precision@100
raw	raw	0.0649	0.0505	0.0440	0.0406	0.0373	0.0353	0.0331	0.0311	0.0292	0.0273
	stopwords	0.0245	0.0212	0.0200	0.0188	0.0176	0.0168	0.0158	0.0154	0.0146	0.0142
	real words	0.0418	0.0345	0.0307	0.0285	0.0264	0.0244	0.0234	0.0223	0.0213	0.0206
keywords	raw	0.0621	0.0536	0.0483	0.0455	0.0431	0.0409	0.0390	0.0374	0.0355	0.0330
	stopwords	0.0256	0.0235	0.0218	0.0203	0.0194	0.0191	0.0181	0.0175	0.0174	0.0173
	real words	0.0416	0.0392	0.0343	0.0316	0.0293	0.0278	0.0271	0.0262	0.0251	0.0242
paraphrase	raw	0.0617	0.0562	0.0502	0.0461	0.0435	0.0412	0.0400	0.0385	0.0366	0.0342
	stopwords	0.0336	0.0303	0.0289	0.0272	0.0265	0.0258	0.0248	0.0235	0.0225	0.0216
	real words	0.0425	0.0365	0.0337	0.0312	0.0295	0.0282	0.0271	0.0260	0.0252	0.0245
real words	raw	0.0254	0.0211	0.0206	0.0195	0.0182	0.0173	0.0165	0.0154	0.0150	0.0141
	stopwords	0.0134	0.0118	0.0114	0.0106	0.0097	0.0094	0.0093	0.0095	0.0092	0.0093
	real words	0.0372	0.0253	0.0213	0.0201	0.0180	0.0170	0.0160	0.0153	0.0144	0.0143

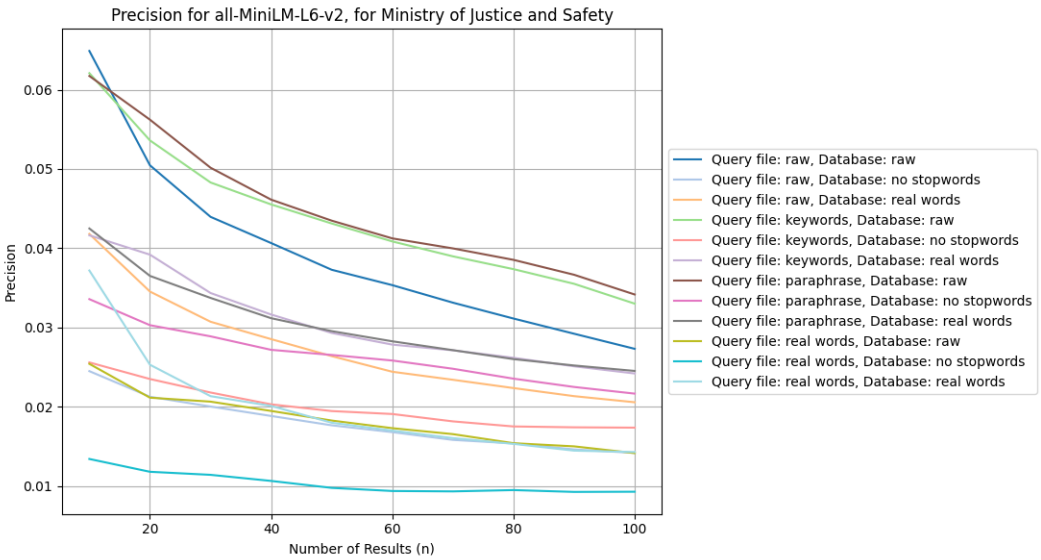

Figure A.44: Dataset: Ministry of Justice and Safety, Model: MiniLM, Metric: Precision

Table A.45: **Recall** at Different Amount of Pages Retrieved Values for **MiniLM** on **Ministry of Justice and Safety**

Query File	Database	Recall@10	Recall@20	Recall@30	Recall@40	Recall@50	Recall@60	Recall@70	Recall@80	Recall@90	Recall@100
raw	raw	0.0125	0.0144	0.0164	0.0179	0.0190	0.0207	0.0217	0.0225	0.0230	0.0234
	stopwords	0.0049	0.0059	0.0076	0.0081	0.0090	0.0094	0.0099	0.0106	0.0110	0.0113
	real words	0.0064	0.0085	0.0099	0.0110	0.0118	0.0136	0.0144	0.0151	0.0159	0.0165
keywords	raw	0.0086	0.0106	0.0124	0.0145	0.0158	0.0173	0.0187	0.0196	0.0208	0.0213
	stopwords	0.0022	0.0034	0.0043	0.0048	0.0054	0.0062	0.0066	0.0070	0.0074	0.0079
	real words	0.0033	0.0058	0.0072	0.0081	0.0091	0.0104	0.0114	0.0123	0.0129	0.0137
paraphrase	raw	0.0070	0.0098	0.0132	0.0146	0.0168	0.0177	0.0199	0.0212	0.0229	0.0235
	stopwords	0.0029	0.0049	0.0078	0.0087	0.0096	0.0108	0.0117	0.0121	0.0141	0.0153
	real words	0.0049	0.0071	0.0085	0.0104	0.0119	0.0128	0.0139	0.0152	0.0162	0.0168
real words	raw	0.0030	0.0043	0.0061	0.0071	0.0084	0.0091	0.0094	0.0096	0.0102	0.0106
	stopwords	0.0008	0.0011	0.0017	0.0022	0.0025	0.0027	0.0034	0.0039	0.0048	0.0052
	real words	0.0077	0.0096	0.0116	0.0127	0.0136	0.0147	0.0154	0.0159	0.0167	0.0171

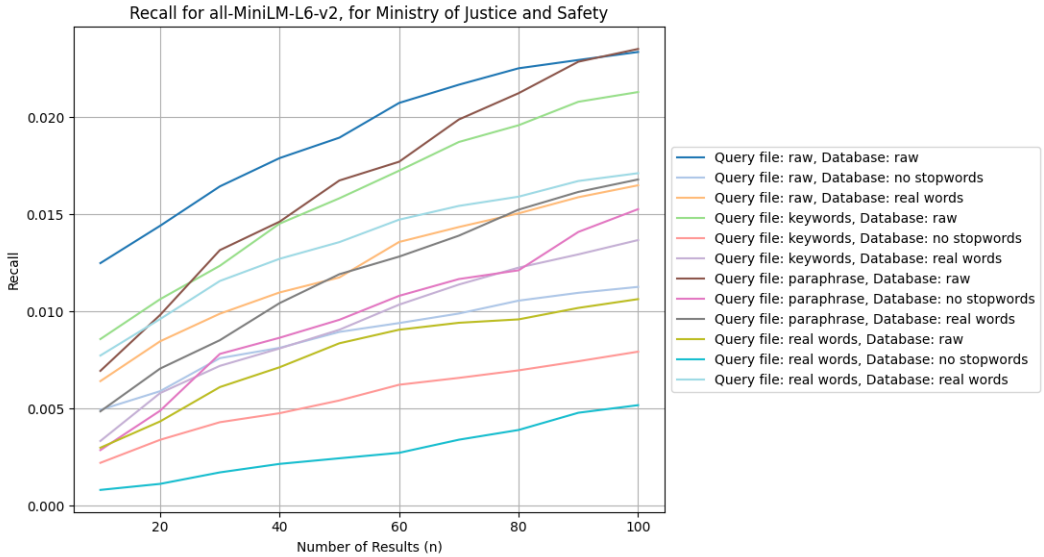

Figure A.45: Dataset: Ministry of Justice and Safety, Model: MiniLM, Metric: Recall

B

Frequency Based Re-evaluation and Weighted Frequency Based Re-evaluation of the Ministries - Full Results

This page is intentionally left blank.

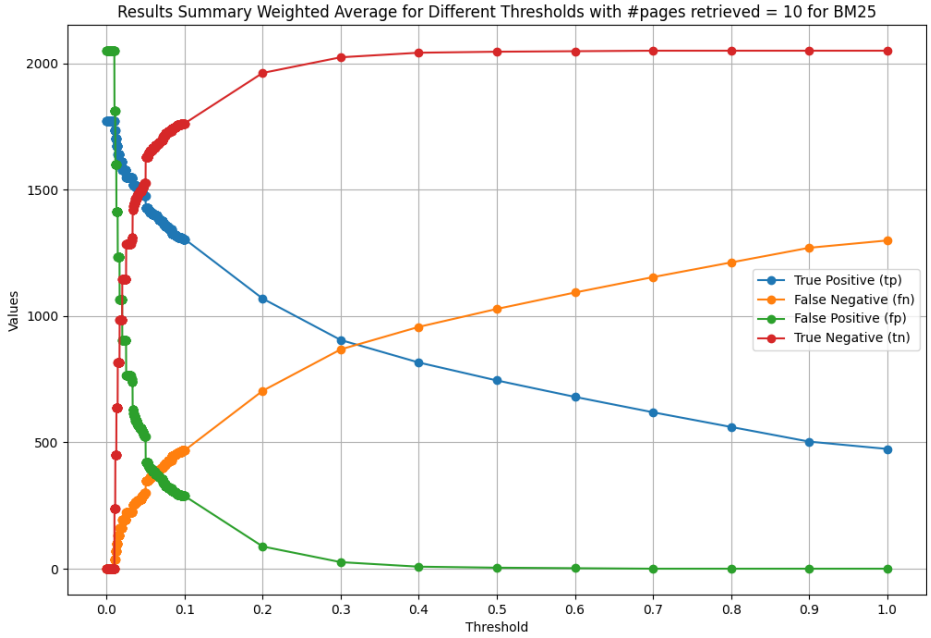

Figure B.1: **Weighted Frequency** for **BM25**, for all ministries averaged, with **n=10**.

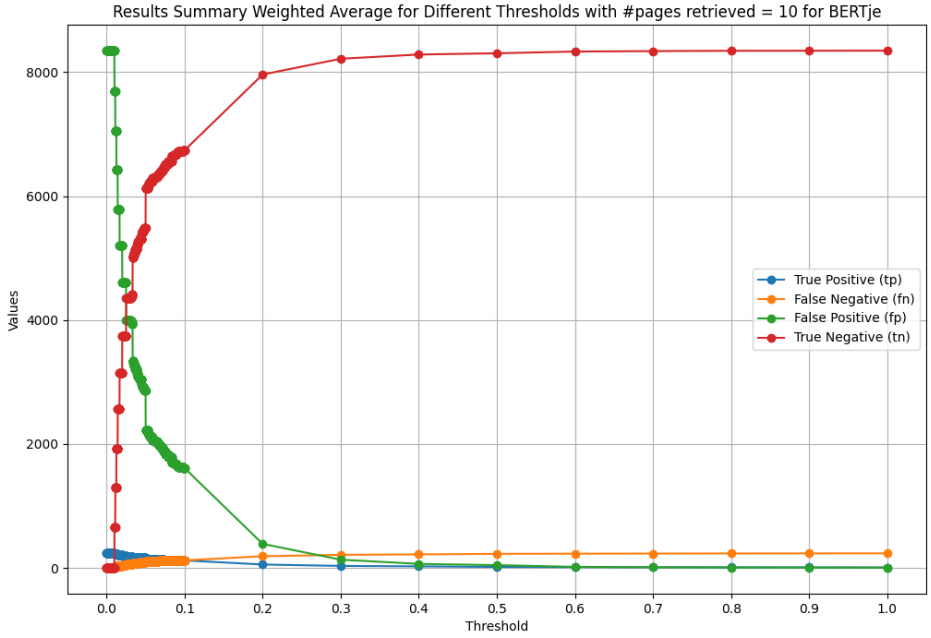

Figure B.2: **Weighted Frequency** for **BERTje**, for all ministries averaged, with **n=10**.

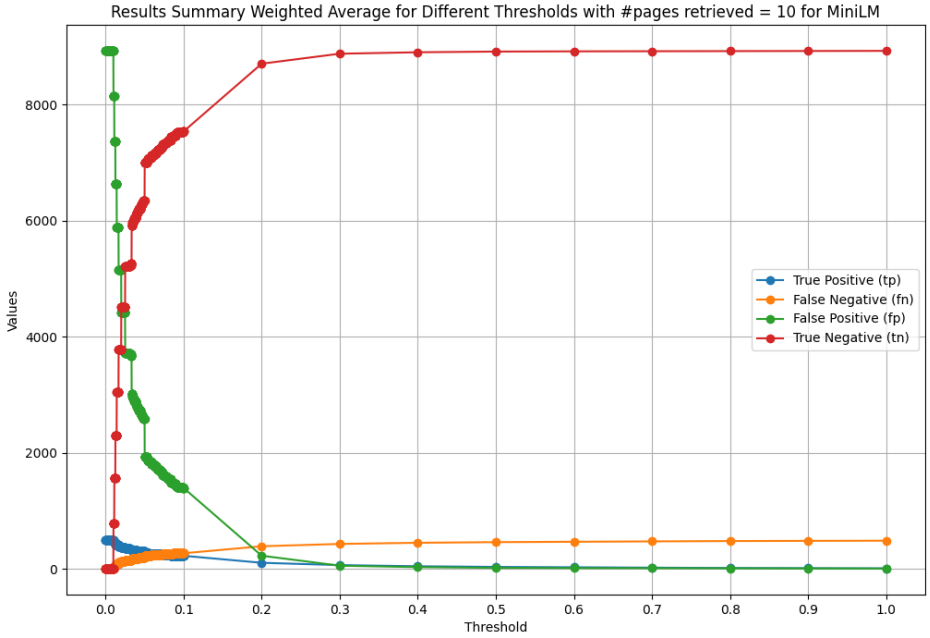

Figure B.3: **Weighted Frequency** for **MiniLM**, for all ministries averaged, with **n=10**.

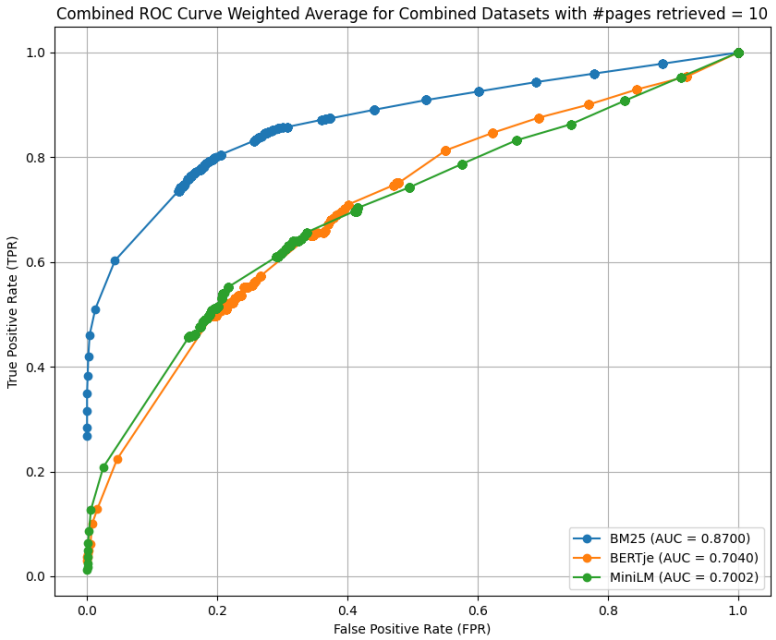

Figure B.4: **ROC Weighted Frequency** for **the different models**, for all ministries averaged with n=10.

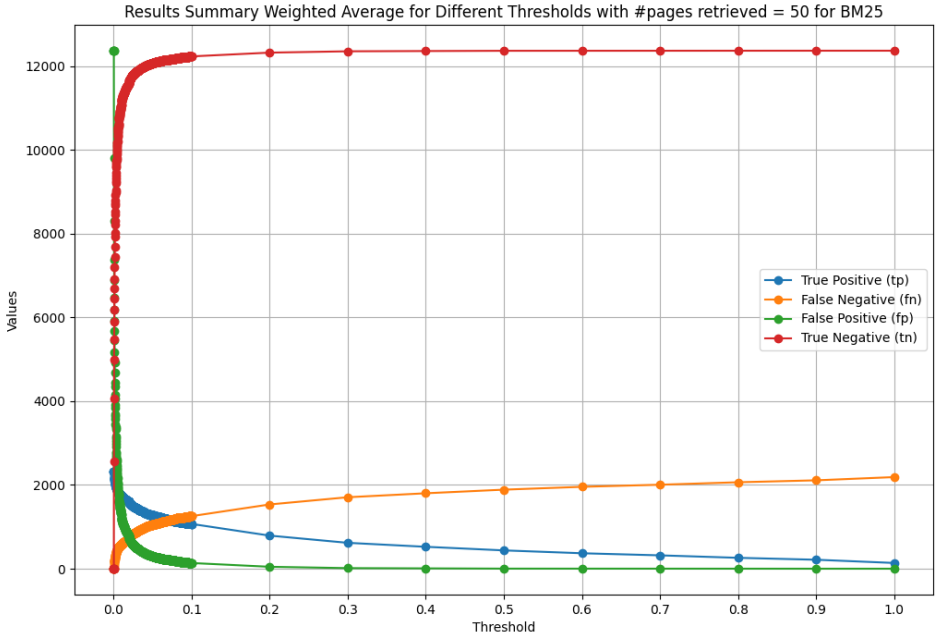

Figure B.5: **Weighted Frequency** for **BM25**, for all ministries averaged, with **n=50**.

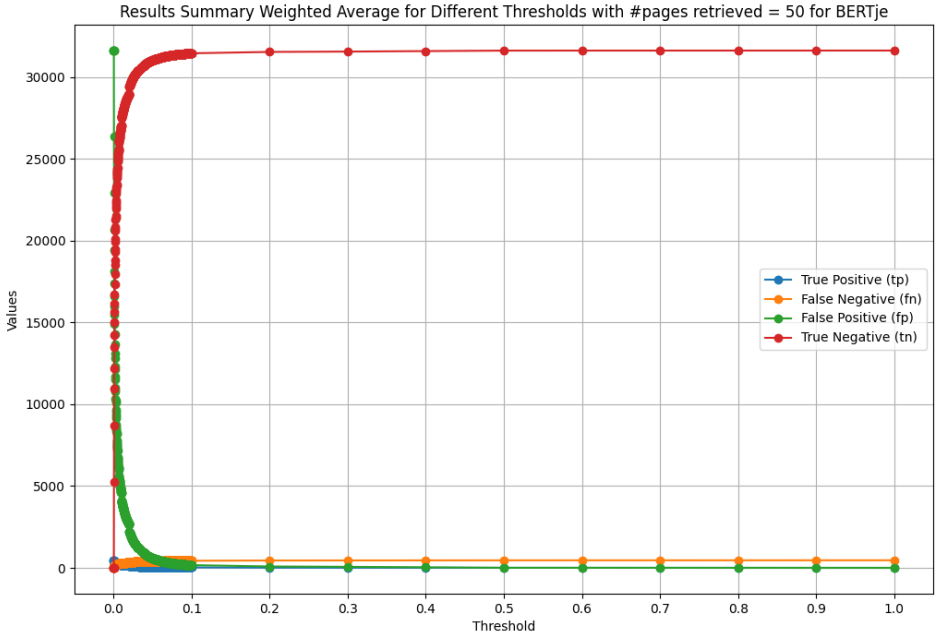

Figure B.6: **Weighted Frequency** for **BERTje**, for all ministries averaged, with **n=50**.

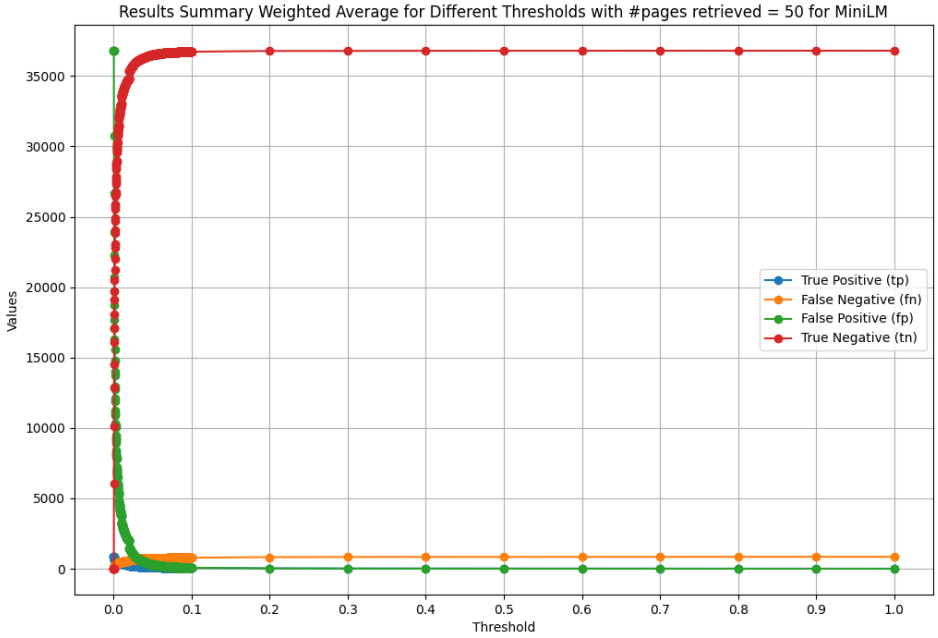

Figure B.7: **Weighted Frequency** for **MiniLM**, for all ministries averaged, with **n=50**.

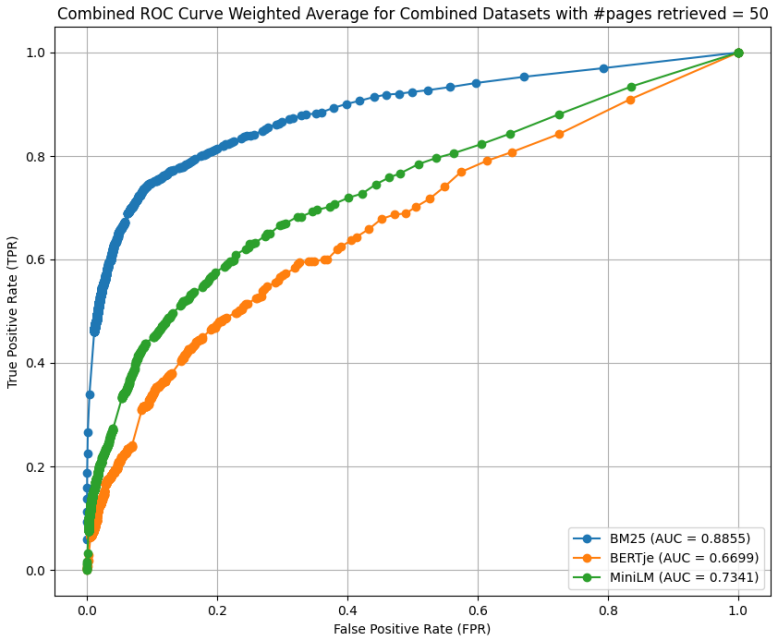

Figure B.8: **ROC Weighted Frequency** for **the different models**, for all ministries averaged with n=50.

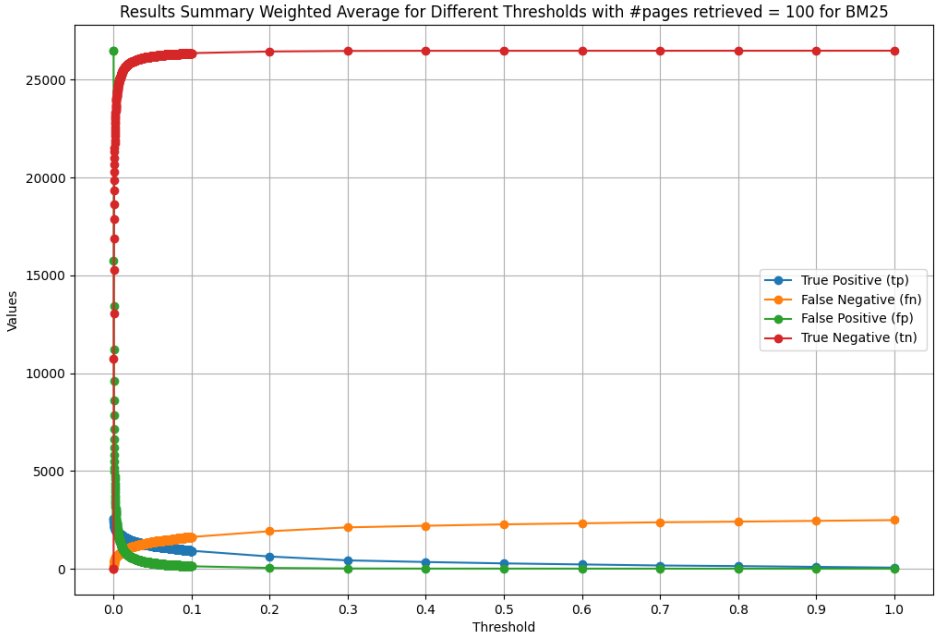

Figure B.9: **Weighted Frequency** for **BM25**, for all ministries averaged, with **n=100**.

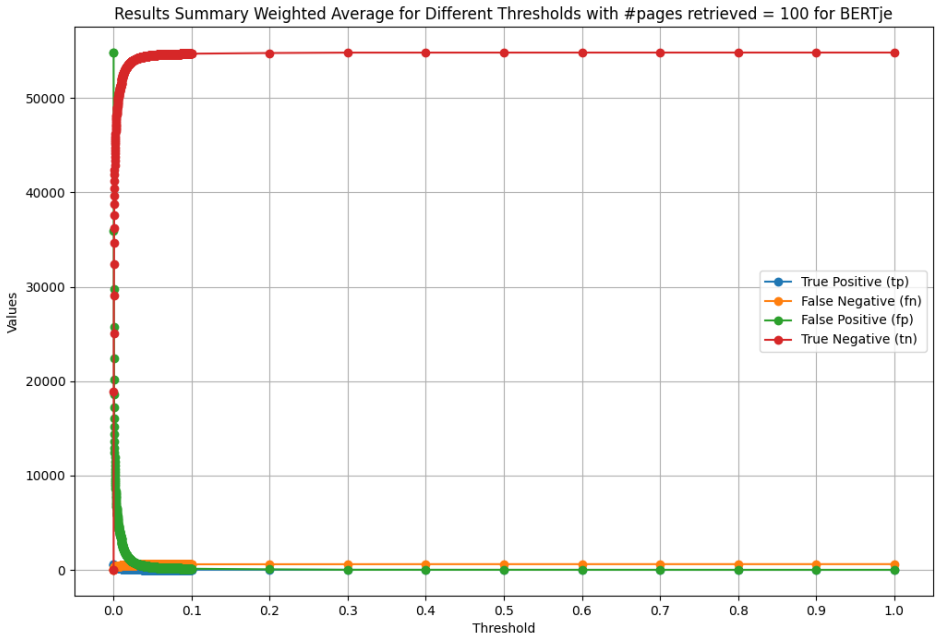

Figure B.10: **Weighted Frequency** for **BERTje**, for all ministries averaged, with **n=100**.

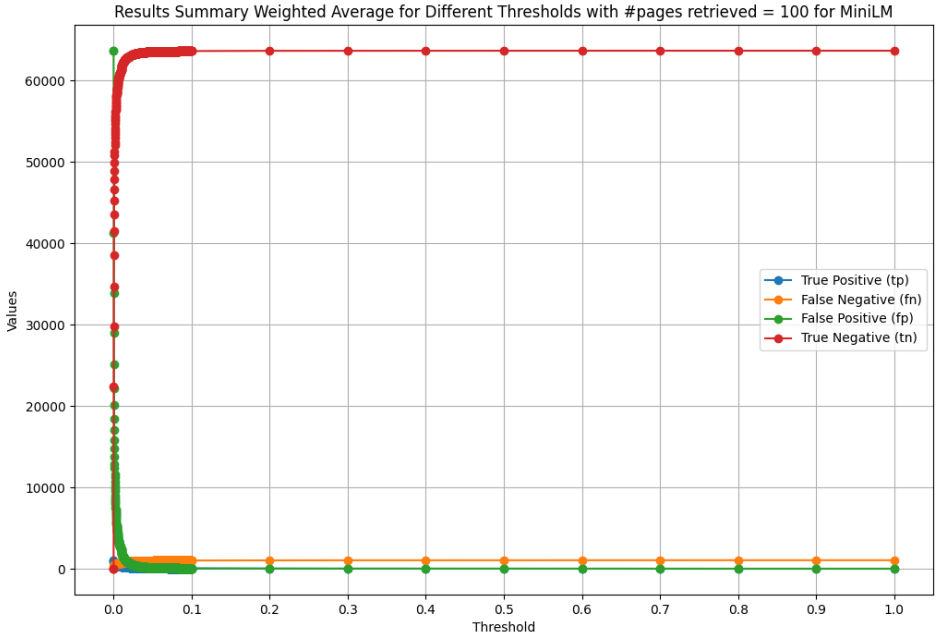

Figure B.11: **Weighted Frequency** for **MiniLM**, for all ministries averaged, with **n=100**.

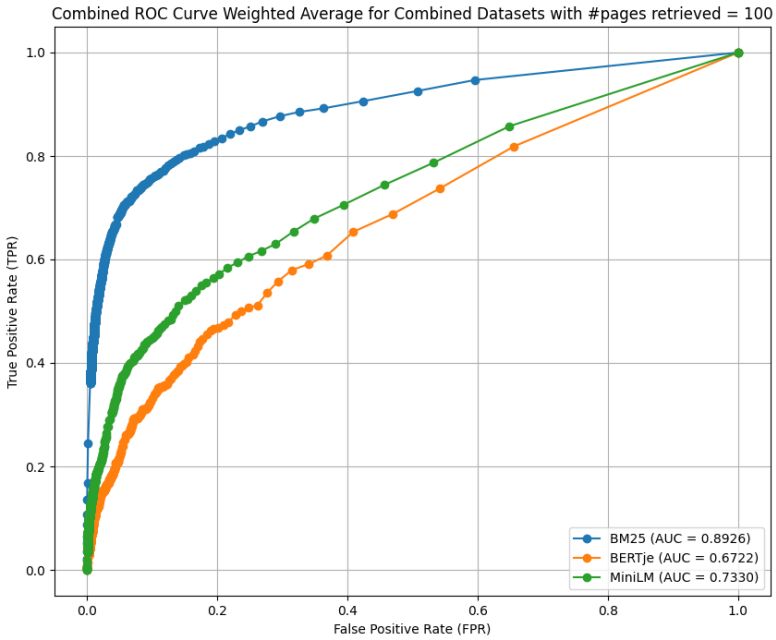

Figure B.12: **ROC Weighted Frequency** for **the different models**, for all ministries averaged with n=100.

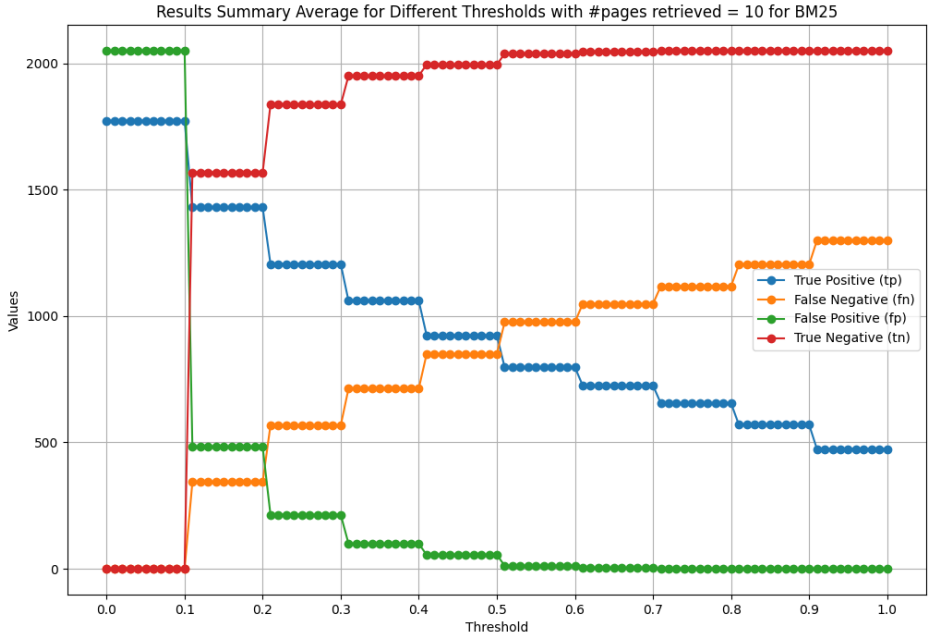

Figure B.13: **Frequency** for **BM25**, for all ministries averaged, with **n=10**.

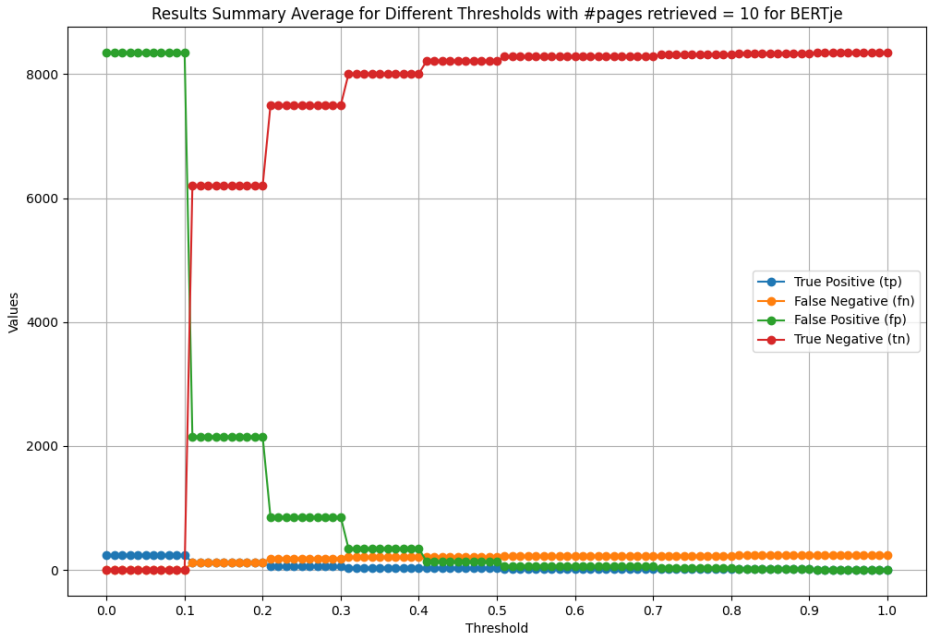

Figure B.14: **Frequency** for **BERTje**, for all ministries averaged, with **n=10**.

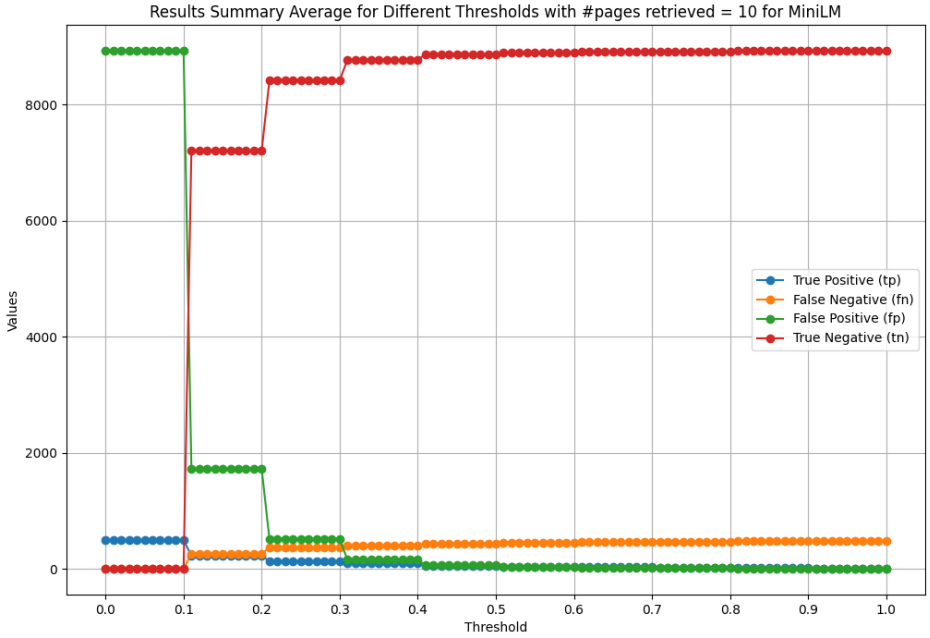

Figure B.15: **Frequency** for **MiniLM**, for all ministries averaged, with **n=10**.

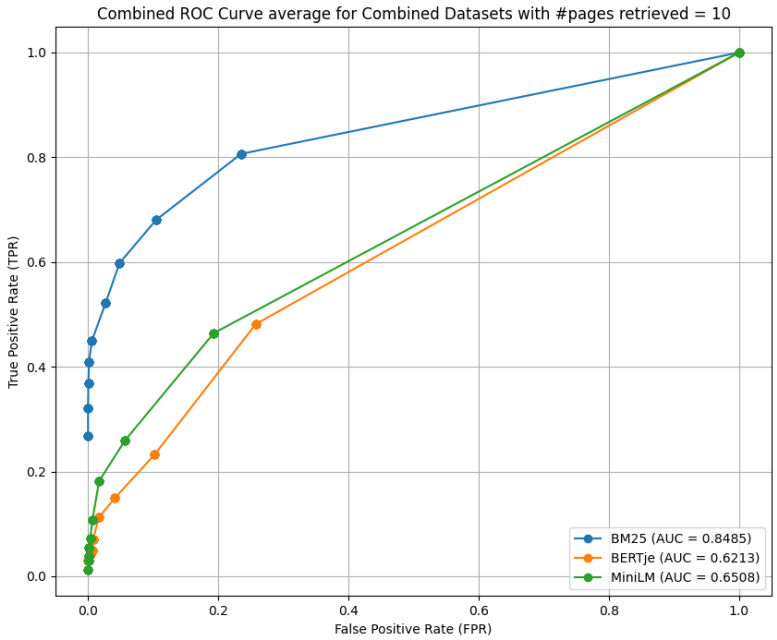

Figure B.16: **ROC Frequency** for **the different models**, for all ministries averaged with n=10.

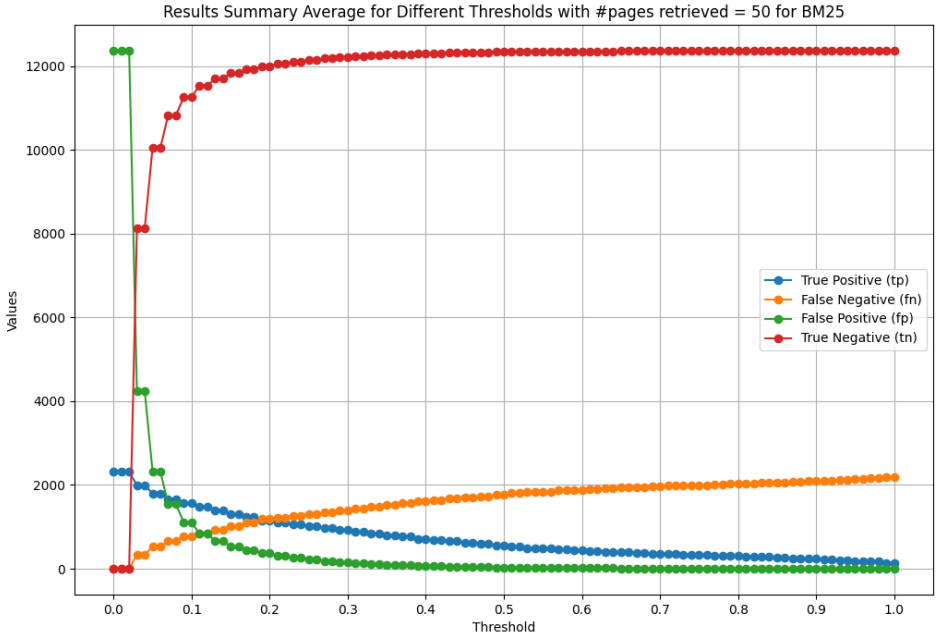

Figure B.17: **Frequency** for **BM25**, for all ministries averaged, with **n=50**.

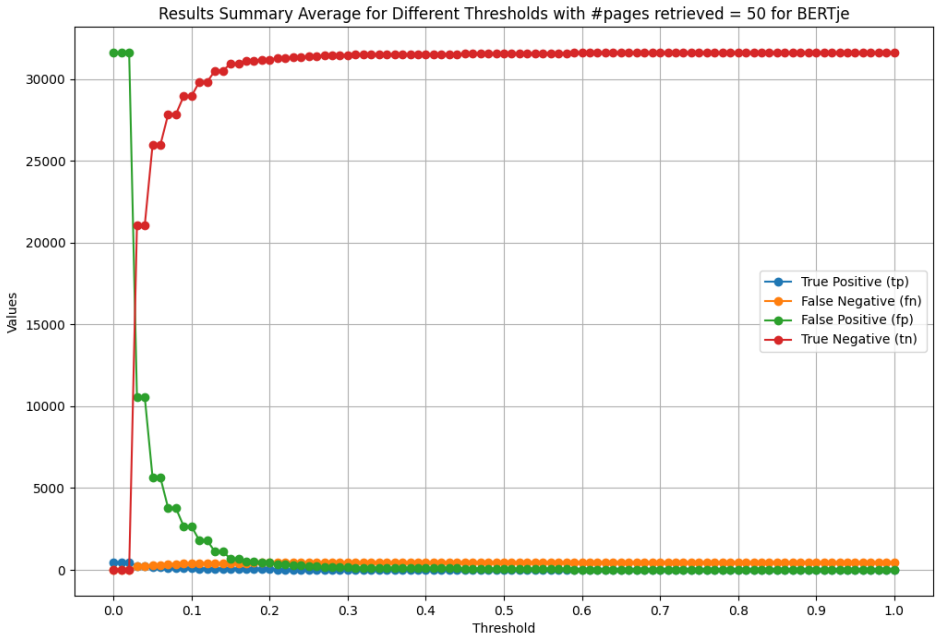

Figure B.18: **Frequency** for **BERTje**, for all ministries averaged, with **n=50**.

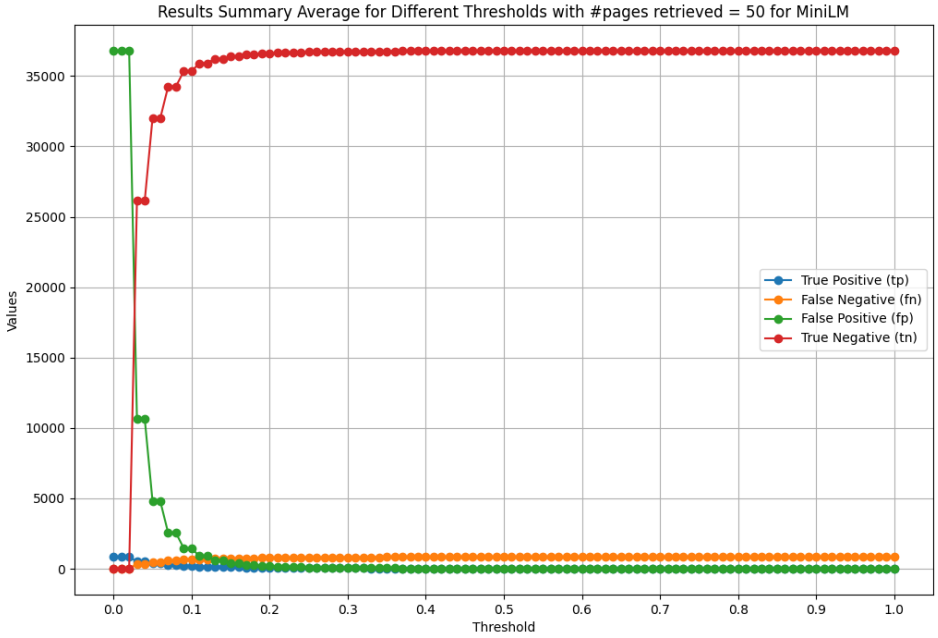

Figure B.19: **Frequency** for **MiniLM**, for all ministries averaged, with **n=50**.

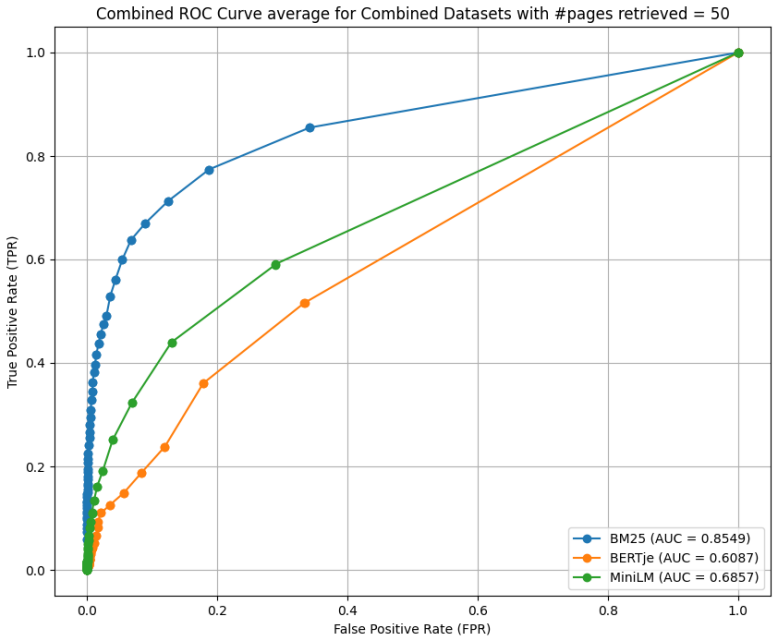

Figure B.20: **ROC Frequency** for **the different models**, for all ministries averaged with n=50.

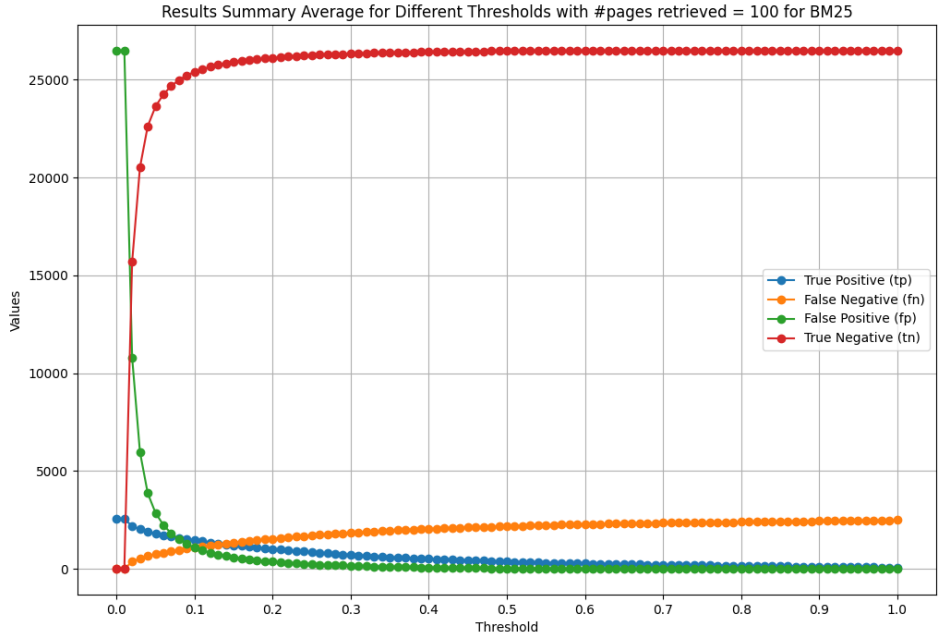

Figure B.21: **Frequency** for **BM25**, for all ministries averaged, with **n=100**.

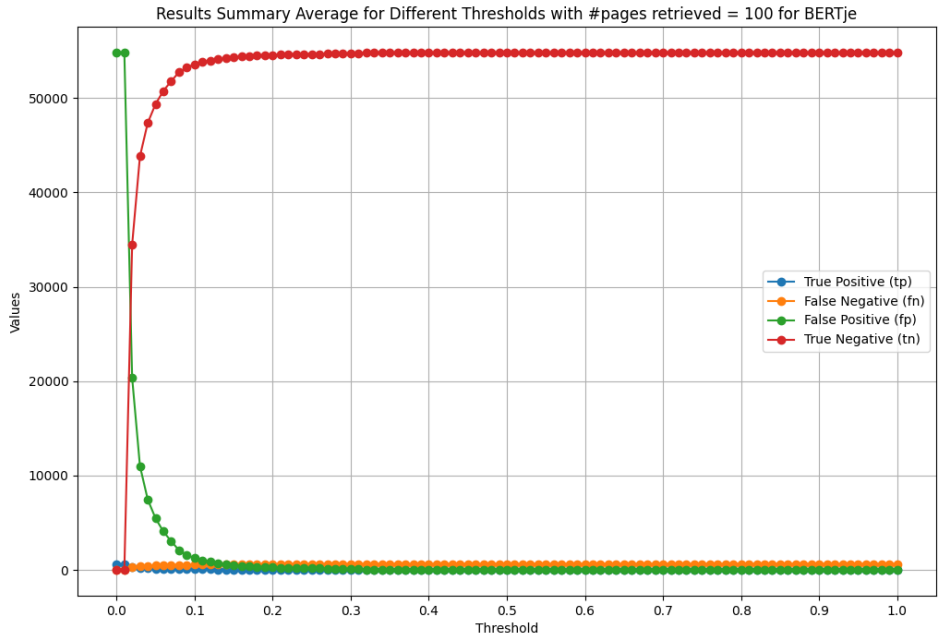

Figure B.22: **Frequency** for **BERTje**, for all ministries averaged, with **n=100**.

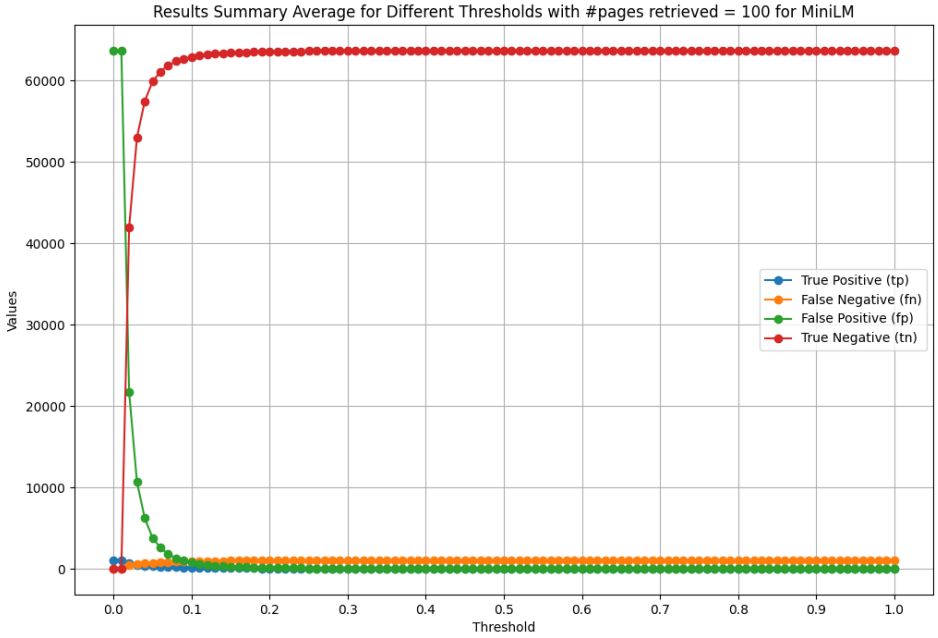

Figure B.23: **Frequency** for **MiniLM**, for all ministries averaged, with **n=100**.

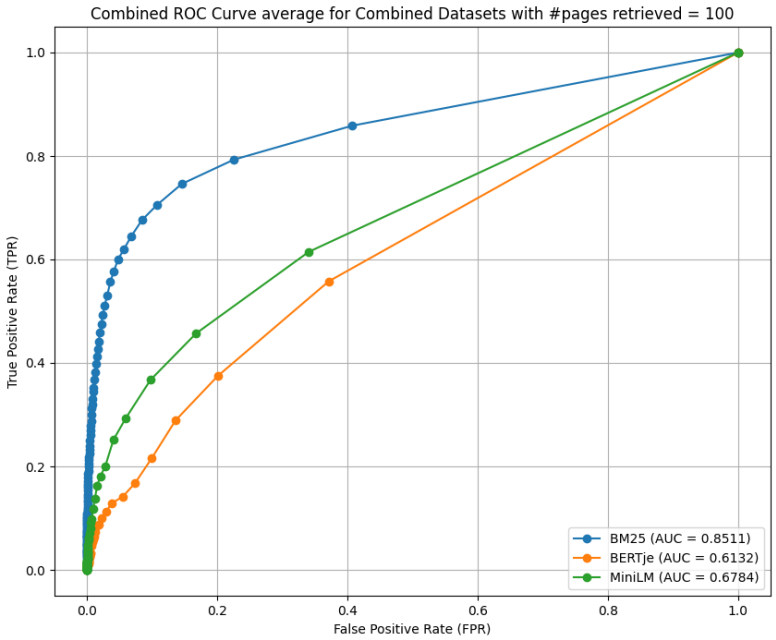

Figure B.24: **ROC Frequency** for **the different models**, for all ministries averaged with n=100.

C

Time Taken per Ministry - Full Results

Process Time (sec)	MinBZK	MinAZ	MinBZ	MinFin	MinJenV
Preprocessing Input					
Whole Request File (no preprocessing)	0	0	0	0	0
Generated Paraphrased Request File	1121	76.7	821	1085	1281
Generated Keywords based on Request File	730.3	66.9	720	882	1067
Real Words in Request File	0.6	0.2	0.7	0.7	0.8
Preprocessing Data					
All Documents (no preprocessing)	0	0	0	0	0
All Documents stopwords removed, stemmed	83.8	3	51.3	112.2	193.1
All Documents, real words	8.8	0.4	4.7	11.7	18.9

Table C.1: Preprocessing Times

Process Time (hh:mm:ss)	MinBZK	MinAZ	MinBZ	MinFin	MinJenV
Ingest BM25S					
All Documents	0:11	0:00	0:06	0:12	0:22
All Documents stopwords removed, stemmed	0:09	0:00	0:05	0:10	0:19
All Documents, real words	0:05	0:00	0:03	0:07	0:12
Ingest BERTje					
All Documents	29:38:24	5:56	8:32:33	33:40:58	128:50:12
All Documents stopwords removed, stemmed	13:10:32	2:56	4:46:32	16:53:45	67:39:16
All Documents, real words	15:36:27	2:42	4:37:54	15:58:03	60:04:42
Ingest MiniLM					
All Documents	27:02:31	4:26	8:11:04	33:16:45	125:46:16
All Documents stopwords removed, stemmed	12:37:26	2:41	4:32:47	17:18:45	66:12:38
All Documents, real words	12:10:46	2:07	4:06:54	15:37:32	58:53:31

Table C.2: Database Creation Times for Different Methods

Process (sec)	Database	MinBZK	MinAZ	MinBZ	MinFin	MinJenV
Evaluatebm25						
Whole Request File	All Documents	57.83	0.59	21.60	79.33	249.73
	All Documents no stopwords, stem	47.05	0.53	24.08	64.13	201.82
	All Documents, real words	44.27	0.52	17.91	59.40	175.56
(Generated) Paraphrased Request File	All Documents	0.59	0.05	0.42	0.80	1.35
	All Documents no stopwords, stem	0.55	0.05	0.45	0.75	1.27
	All Documents, real words	0.45	0.05	0.41	0.66	1.04
(Generated) Keywords on Request File	All Documents	0.54	0.05	0.40	0.80	1.29
	All Documents no stopwords, stem	0.50	0.05	0.43	0.74	1.18
	All Documents, real words	0.42	0.05	0.38	0.67	0.97
Real Words in Request File	All Documents	35.60	0.36	12.28	48.66	141.31
	All Documents no stopwords, stem	28.82	0.31	10.08	40.00	115.61
	All Documents, real words	32.35	0.35	11.00	43.93	128.48
Evaluate BERTje						
Whole Request File	All Documents	81.79	9.98	27.84	109.58	106.97
	All Documents no stopwords, stem	49.04	9.30	27.68	112.32	105.65
	All Documents, real words	51.54	8.84	27.75	115.03	105.92
(Generated) Paraphrased Request File	All Documents	8.24	6.70	5.22	10.48	9.78
	All Documents no stopwords, stem	5.99	6.15	4.85	8.96	9.07
	All Documents, real words	6.38	5.91	6.09	9.70	9.33
(Generated) Keywords on Request File	All Documents	7.49	8.96	5.22	10.63	9.78
	All Documents no stopwords, stem	6.01	5.97	4.78	8.48	9.60
	All Documents, real words	6.67	6.37	5.26	9.25	9.30
Real Words in Request File	All Documents	27.23	7.27	14.55	55.75	53.16
	All Documents no stopwords, stem	26.24	7.29	15.26	54.92	53.12
	All Documents, real words	27.38	3.66	14.65	56.77	53.13
Evaluate MiniLM						
Whole Request File	All Documents	98.47	5.16	31.42	126.05	127.19
	All Documents no stopwords, stem	96.98	5.18	30.16	103.94	124.78
	All Documents, real words	97.49	5.15	30.16	71.55	125.76
(Generated) Paraphrased Request File	All Documents	6.95	3.26	4.96	8.99	8.60
	All Documents no stopwords, stem	5.71	3.20	4.71	6.51	7.74
	All Documents, real words	6.31	3.47	5.01	6.75	7.85
(Generated) Keywords on Request File	All Documents	6.39	3.29	4.89	7.86	8.07
	All Documents no stopwords, stem	5.77	3.20	4.69	6.56	7.54
	All Documents, real words	5.79	3.38	4.87	6.38	7.66
Real Words in Request File	All Documents	52.24	4.36	15.98	67.35	66.94
	All Documents no stopwords, stem	51.23	4.15	16.01	39.72	66.01
	All Documents, real words	51.68	4.27	16.36	40.15	66.49

Table C.3: Evaluation Times for Different Models for the Ministries Separately

Bibliography

[1] *Afhandeling informatieverzoeken Wet open overheid door ministeries verder vertraagd*. 2024. URL: https://openstate.eu/nl/2024/02/afhandeling-informatieverzoeken-wet-open-overheid-door-ministeries-verder-vertraagd/.

[2] Adam Berger et al. "Bridging the lexical chasm: statistical approaches to answer-finding". In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '00. Athens, Greece: Association for Computing Machinery, 2000, pp. 192–199. ISBN: 1581132263. DOI: 10.1145/345508.345576. URL: https://doi.org/10.1145/345508.345576.

[3] Jose Camacho-Collados and Mohammad Taher Pilehvar. *On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis*. 2018. arXiv: 1707.01780 [cs.CL].

[4] Stefano Ceri et al. "An Introduction to Information Retrieval". In: *Web Information Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 3–11. ISBN: 978-3-642-39314-3. DOI: 10.1007/978-3-642-39314-3_1. URL: https://doi.org/10.1007/978-3-642-39314-3_1.

[5] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[6] Guido Enthoven et al. *Ondraaglijk traag Analyse afhandeling Wob-verzoeken*. Jan. 2022. URL: https://openstate.eu/wp-content/uploads/sites/14/2022/01/Ondraaglijk-traag-280122-def.pdf.

[7] H. Ferhatosmanoglu et al. "Approximate nearest neighbor searching in multimedia databases". In: *Proceedings 17th International Conference on Data Engineering*. 2001, pp. 503–511. DOI: 10.1109/ICDE.2001.914864.

[8] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. *SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking*. 2021. arXiv: 2107.05720 [cs.IR].

[9] Luyu Gao et al. *Precise Zero-Shot Dense Retrieval without Relevance Labels*. 2022. arXiv: 2212.10496 [cs.IR]. URL: https://arxiv.org/abs/2212.10496.

[10] Yunfan Gao et al. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024. arXiv: 2312.10997 [cs.CL]. URL: https://arxiv.org/abs/2312.10997.

[11] Koen Giezeman and Diesfeldt Dominique. *FAQ: Er is een Woo-verzoek over mij ingediend, wat nu?* May 2023. URL: https://www.rijksoverheid.nl/documenten/rapporten/2022/07/18/overzicht-per-ministerie-van-wob-woo-verzoeken-in-behandeling.

[12] David Grangier, Alessandro Vinciarelli, and Hervé Bourlard. "Information Retrieval on Noisy Text". In: (2003).

[13] Maura R. Grossman, Gordon V. Cormack, and Jason R. Baron. "Does the LLMperor Have New Clothes? Some Thoughts on the Use of LLMs in eDiscovery". In: *SSRN* (2024). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4949879.

[14] Kelvin Guu et al. *REALM: Retrieval-Augmented Language Model Pre-Training*. 2020. arXiv: 2002.08909 [cs.CL].

[15] Yikun Han, Chunjiang Liu, and Pengfei Wang. *A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge*. 2023. arXiv: 2310.11703 [cs.DB].

[16] Sebastian Hofstätter et al. *Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling*. 2021. arXiv: 2104.06967 [cs.IR].

[17] *Introducing Contextual Retrieval*. Sept. 2024. URL: https://www.anthropic.com/news/contextual-retrieval.

[18] Gautier Izacard et al. *Unsupervised Dense Information Retrieval with Contrastive Learning*. 2022. arXiv: 2112.09118 [cs.IR].

[19] Zhengbao Jiang et al. "X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 5943–5959. DOI: 10.18653/v1/2020.emnlp-main.479. URL: https://aclanthology.org/2020.emnlp-main.479.

[20] Jeff Johnson, Matthijs Douze, and Hervé Jégou. *Billion-scale similarity search with GPUs*. 2017. arXiv: 1702.08734 [cs.CV].

[21] *Kabinet neemt maatregelen voor een betere uitvoering en uitvoerbaarheid van de Wet open overheid*. June 2024. URL: https://www.rijksoverheid.nl/actueel/nieuws/2024/06/21/kabinet-neemt-maatregelen-voor-een-betere-uitvoering-en-uitvoerbaarheid-van-de-wet-open-overheid#:~:text=Het%20kabinet%20neemt%20maatregelen%20om,tussen%20overheidsorganisaties%20en%20Woo%2Dverzoekers.

[22] Ammar Ismael Kadhim. "Term Weighting for Feature Extraction on Twitter: A Comparison Between BM25 and TF-IDF". In: *2019 International Conference on Advanced Science and Engineering (ICOASE)*. 2019, pp. 124–128. DOI: 10.1109/ICOASE.2019.8723825.

[23] Vladimir Karpukhin et al. *Dense Passage Retrieval for Open-Domain Question Answering*. 2020. arXiv: 2004.04906 [cs.CL].

[24] Enkelejda Kasneci et al. "ChatGPT for good? On opportunities and challenges of large language models for education". In: *Learning and Individual Differences* 103 (2023), p. 102274. ISSN: 1041-6080. DOI: https://doi.org/10.1016/j.lindif.2023.102274. URL: https://www.sciencedirect.com/science/article/pii/S1041608023000195.

[25] Edda Leopold and Jörg Kindermann. "Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?" In: *Machine Learning* 46 (Jan. 2002), pp. 423–444. DOI: 10.1023/A:1012491419635.

[26] Patrick Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[27] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. "In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval". In: *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*. Ed. by Anna Rogers et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 163–173. DOI: 10.18653/v1/2021.repl4nlp-1.17. URL: https://aclanthology.org/2021.repl4nlp-1.17.

[28] Edward Loper and Steven Bird. *NLTK: The Natural Language Toolkit*. 2002. arXiv: cs/0205028 [cs.CL]. URL: https://arxiv.org/abs/cs/0205028.

[29] Yi Luan et al. "Sparse, Dense, and Attentional Representations for Text Retrieval". In: *Transactions of the Association for Computational Linguistics* 9 (Apr. 2021), pp. 329–345. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00369. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00369/1924040/tacl_a_00369.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00369.

[30] Yuanhua Lv and ChengXiang Zhai. "Lower-bounding term frequency normalization". In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. CIKM '11. Glasgow, Scotland, UK: Association for Computing Machinery, 2011, pp. 7–16. ISBN: 9781450307178. DOI: 10.1145/2063576.2063584. URL: https://doi.org/10.1145/2063576.2063584.

[31] Varun Magesh et al. *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*. 2024. arXiv: 2405.20362 [cs.CL]. URL: https://arxiv.org/abs/2405.20362.

[32] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[33] Bhaskar Mitra and Nick Craswell. "An Introduction to Neural Information Retrieval". In: *Foundations and Trendső in Information Retrieval* 13.1 (2018), pp. 1–126. ISSN: 1554-0669. DOI: 10.1561/1500000061. URL: http://dx.doi.org/10.1561/1500000061.

[34] Mandar Mitra and BB Chaudhuri. "Information retrieval from documents: A survey". In: *Information retrieval* 2 (2000), pp. 141–163.

[35] Niklas Muennighoff et al. *MTEB: Massive Text Embedding Benchmark*. 2023. arXiv: 2210.07316 [cs.CL]. URL: https://arxiv.org/abs/2210.07316.

[36] Karthik Narasimhan, Adam Yala, and Regina Barzilay. "Improving Information Extraction by Acquiring External Evidence with Reinforcement Learning". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by Jian Su, Kevin Duh, and Xavier Carreras. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2355–2365. DOI: 10.18653/v1/D16-1261. URL: https://aclanthology.org/D16-1261.

[37] *Nog altijd veel mis met Wet open overheid: 'Het wordt als corveetaak gezien'*. 2024. URL: https://nos.nl/nieuwsuur/artikel/2509029-nog-altijd-veel-mis-met-wet-open-overheid-het-wordt-als-corveetaak-gezien.

[38] Rodrigo Nogueira and Kyunghyun Cho. *Passage Re-ranking with BERT*. 2020. arXiv: 1901.04085 [cs.IR].

[39] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774.

[40] *Opvragen overheidsinformatie (Woo-verzoek)*. Mar. 2024. URL: https://www.rijksoverheid.nl/wetten-en-regelingen/productbeschrijvingen/opvragen-overheidsinformatie-woo-verzoek.

[41] *Overzicht per ministerie van Wob/Woo-verzoeken in behandeling*. July 2022. URL: https://www.stibbe.com/nl/publications-and-insights/faq-er-is-een-woo-verzoek-over-mij-ingediend-wat-nu.

[42] *Overzicht Woo-verzoeken met verbeurde dwangsommen uitvraag januari 2024*. Jan. 2024. URL: https://app.1848.nl/document/tkapi/19050.

[43] Hugo van der Parre. *Meerderheid Woo-Verzoeken Bij ministeries is ver over tijd*. July 2022. URL: https://www.vvoj.org/2022/07/20/meerderheid-woo-verzoeken-bij-ministeries-is-ver-over-tijd/.

[44] Telmo Pires, Eva Schlinger, and Dan Garrette. *How multilingual is Multilingual BERT?* 2019. arXiv: 1906.01502 [cs.CL].

[45] Wisam Qader, Musa M. Ameen, and Bilal Ahmed. "An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges". In: June 2019, pp. 200–204. DOI: 10.1109/IEC47844.2019.8950616.

[46] Zahra Rahimi and Mohammad Mehdi Homayounpour. "The impact of preprocessing on word embedding quality: a comparative study". In: *Lang. Resour. Eval.* 57.1 (Oct. 2022), pp. 257–291. ISSN: 1574-020X. DOI: 10.1007/s10579-022-09620-5. URL: https://doi.org/10.1007/s10579-022-09620-5.

[47] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 2019. arXiv: 1908.10084 [cs.CL].

[48] Stephen Robertson and Hugo Zaragoza. "The Probabilistic Relevance Framework: BM25 and Beyond". In: *Foundations and Trends in Information Retrieval* 3 (Jan. 2009), pp. 333–389. DOI: 10.1561/1500000019.

[49] Jacques Savoy and Nada Naji. "Comparative information retrieval evaluation for scanned documents". In: July 2011, pp. 527–534.

[50] Karen Sparck Jones. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". In: *Journal of Documentation* 28.1 (Jan. 1972), pp. 11–21. ISSN: 0022-0418. DOI: 10.1108/eb026526. URL: https://doi.org/10.1108/eb026526.

[51] Gemini Team et al. *Gemini: A Family of Highly Capable Multimodal Models*. 2024. arXiv: 2312.11805 [cs.CL]. URL: https://arxiv.org/abs/2312.11805.

[52] Nandan Thakur et al. *BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models*. 2021. arXiv: 2104.08663 [cs.IR].

[53] Ravi Theja. *Evaluating the Ideal Chunk Size for a RAG System using LlamaIndex*. Oct. 2023. URL: https://www.llamaindex.ai/blog/evaluating-the-ideal-chunk-size-for-a-rag-system-using-llamaindex-6207e5d3fec5.

[54] Andrew Trotman, Antti Puurula, and Blake Burgess. "Improvements to BM25 and Language Models Examined". In: *Proceedings of the 19th Australasian Document Computing Symposium*. ADCS '14. Melbourne, VIC, Australia: Association for Computing Machinery, 2014, pp. 58–65. ISBN: 9781450330008. DOI: 10.1145/2682862.2682863. URL: https://doi.org/10.1145/2682862.2682863.

[55] Alper Kürat Uysal and Serkan Gunal. "The impact of preprocessing on text classification". In: *Information Processing & Management* 50 (Jan. 2014), pp. 104–112. DOI: 10.1016/j.ipm.2013.08.006.

[56] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[57] Wietse de Vries et al. *BERTje: A Dutch BERT Model*. 2019. arXiv: 1912.09582 [cs.CL].

[58] Wenhui Wang et al. *MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers*. 2020. arXiv: 2002.10957 [cs.CL]. URL: https://arxiv.org/abs/2002.10957.

[59] *Wat Gebeurt Er Nadat ik een woo-verzoek heb gedaan?* Feb. 2024. URL: https://www.rijksoverheid.nl/onderwerpen/wet-open-overheid-woo/vraag-en-antwoord/wat-gebeurt-er-nadat-ik-een-woo-verzoek-heb-ingediend.

[60] "What is Information Retrieval?" In: (). URL: https://www.elastic.co/what-is/information-retrieval.

[61] John S. Whissell and Charles L. A. Clarke. "Improving document clustering using Okapi BM25 feature weighting". In: *Inf. Retr.* 14.5 (Oct. 2011), pp. 466–487. ISSN: 1386-4564. DOI: 10.1007/s10791-011-9163-y. URL: https://doi.org/10.1007/s10791-011-9163-y.

[62] Lee Xiong et al. *Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval*. 2020. arXiv: 2007.00808 [cs.IR].

[63] Wei Yang et al. “Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models”. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'19. Paris, France: Association for Computing Machinery, 2019, pp. 1129–1132. ISBN: 9781450361729. DOI: 10.1145/3331184.3331340. URL: https://doi.org/10.1145/3331184.3331340.

[64] Jingtao Zhan et al. *Learning To Retrieve: How to Train a Dense Retrieval Model Effectively and Efficiently*. 2020. arXiv: 2010.10469 [cs.IR].

[65] Jingtao Zhan et al. *Optimizing Dense Retrieval Model Training with Hard Negatives*. 2021. arXiv: 2104.08051 [cs.IR]. URL: https://arxiv.org/abs/2104.08051.

[66] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. “Understanding bag-of-words model: a statistical framework”. In: *International Journal of Machine Learning and Cybernetics* 1.1 (Dec. 2010), pp. 43–52. ISSN: 1868-808X. DOI: 10.1007/s13042-010-0001-0. URL: https://doi.org/10.1007/s13042-010-0001-0.

[67] *Zoek & Vind: Alles op een plek*. July 2022. URL: https://www.informatiehuishouding.nl/binaries/informatiehuishouding/documenten/publicaties/2022/07/06/facsheet-zoek--vind-alles-op-een-plek/Zoek+%26+Vind+Alles+op+een+plek.pdf.