A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

INCM & NovaSBE PBL Project – Retirement Law Query System – Performance Analysis & Evaluation of BERT for Extractive Question Answering of Portuguese Retirement Law

Business Analytics, 2021/2022

NOVA School of Business and Economics

Ricardo Araújo

Work project carried out under the supervision of:

Patrícia Xufre

João Magalhães (co-advisor)

17/12/2021

Abstract

This study presents a performance evaluation of the model BERT for an Extractive Question Answering task applied to the domain of Portuguese Retirement Law. By using multiple standard metrics in information retrieval solutions, the study measures how the Portuguese BERT performs as a semantic search solution on Portuguese Retirement Law by comparing its answers with a testing set composed of 180 questions and answers provided by INCM. The study revealed that the solution showed positive results. A form was also distributed to INCM, which indicated positive feedback when comparing it with the current solution implemented. This project showed that the Portuguese BERT is a very promising tool that can be used for Extractive Question Answering and for semantic search on various other domains.

Keywords: Business Analytics, Data Science, Deep Learning, BERT, Semantic Search, Question Answering

Acknowledgements

The present study could not be possible without the help and efforts of my fellow colleagues and friends: Angelo Figueiral, Diogo Cardoso and Pedro Rolim. This thesis could not have been completed without the expertise and guidance of Patrícia Xufre and João Magalhães. A debt of gratitude is also owed to Deepset.AI for their support on any issue raised. Finally, I would like to thank my parents, brothers, and close friends for the endless support.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Table of Contents

Introduction

The *Imprensa Nacional – Casa da Moeda* (INCM) is the Portuguese Mint and Official Printing Office located in Lisbon, Portugal, and is a public company fully owned by the Portuguese Government and administratively subordinated to the Portuguese Ministry of Finance. INCM is responsible for the minting of coins, the production of identity cards, including the Portuguese citizenship card, passports, and drivers' licenses, as well as for the official publications, also on amendments to the law, which are assigned to the Portuguese Official Gazette (Diário da República - DRE) and published in its official site (dre.pt).

Through the incorporation of new technologies into its daily business activities, INCM has a clear focus on the future. It was in that context that the need for a modernization of Law consultation came into existence. Through the "Decoding Legislation for Citizens" project, INCM intends to study the feasibility of implementing a chatbot solution to query the Portuguese Law for Portuguese citizens. Specifically, the use case selected for the project was the Retirement Law Statute, since this is one of the most queried subjects by online users, while being one of the most difficult to obtain answers to.

The problem with the current solution can be encapsulated in the following example. On one hand, a vast number of searches performed on dre.pt were made using the keyword "reforma". On the other hand, most of the law text contains the keyword "aposentação". This is just an example that represents a massive roadblock, since there are a variety of ways the user can make the same question and the current solution cannot find the correct answer for most of them. Moreover, since the currently implemented search engine is **keyword-based**, slight orthographic errors are enough for it to not return any results.

Figure 1 shows the difference between a keyword-based search engine and a semantic search one.

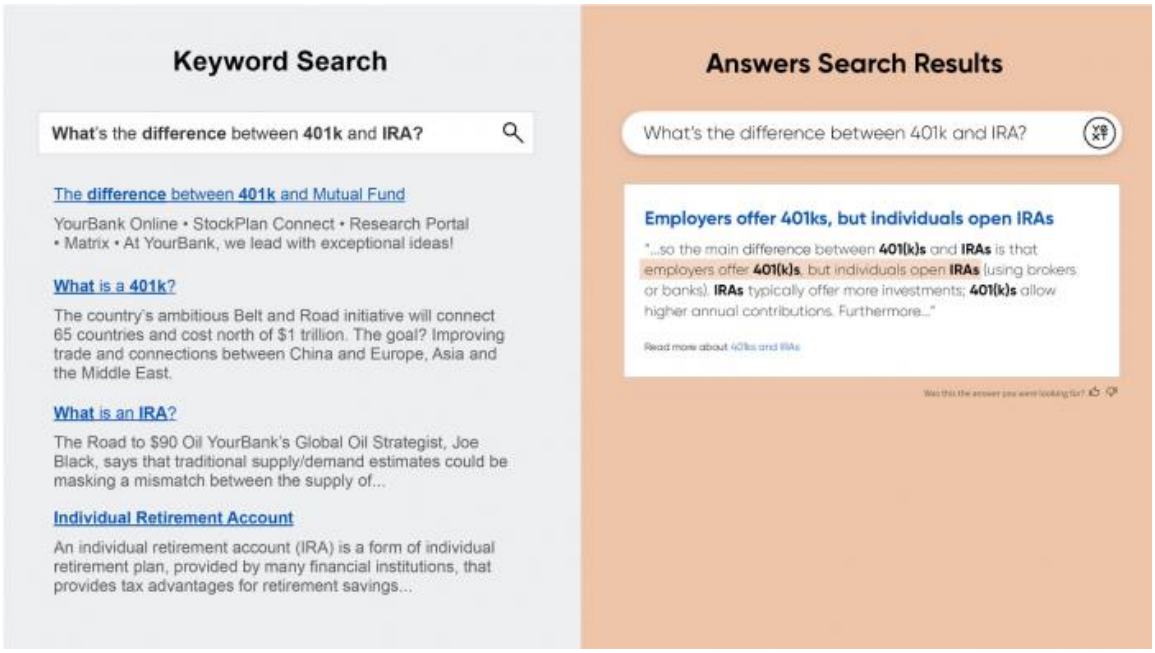


Figure 1 - Keyword-based search engine (left) vs a semantic search engine which tries to understand the intent and context of the search (right).

Source: https://www.yext.com/blog/2021/02/extractive-qa-the-next-big-thing-in-search/ accessed in December 2021

By providing an advanced semantic search engine solution, INCM would be able to provide superior service to their users, since it would offer a way of querying the collection of Portuguese Laws' texts using non-technical Law terms.

INCM envisioned a solution which enabled one to query the Law texts by writing as if one was chatting with another person. The answers should be returned without any text simplification and within their Law context. Furthermore, any dependency, such as supporting legislation and their URL, should also be returned. Another requirement from INCM was that answers to the questions should be objective and the solution should also be extractive only, i.e. the model would have to find and return the right answer to the user exactly as it is written in the article. In other words, translation of law concepts to a more accessible language **must not** occur.

The ‘Data Curation’ Section describes the dataset of the project and all the steps to prepare the data to be inputted in the model. The ‘Exploratory Data Analysis’ (EDA) Section describes the analysis and experimentation performed on the data to better understand the dataset. The ‘Bidirectional Encoder Representations from Transformers’ (BERT) and ‘Haystack’ Sections present the tools, describing how they are used in this specific project and the synergies they create together. The ‘Performance Analysis of the Solution’ Section contains the methodology, explanation of the choices made, evaluation of the results and comparison with INCM’s current solution. The ‘Limitations’ Section describes the limitations and obstacles encountered. It is followed by the ‘Conclusion’ Section, which summarizes all the results of the study. The ‘Further Work’ Section presents the relevant further work that can be done and what can be added to increase the performance of the solution.

The end goal of the thesis is to build and evaluate the quality of an accessible solution which, triggered by a question, retrieves the relevant information from the Portuguese Retirement Law as an answer. To quantitatively measure the performance of the new state-of-the-art model used in this project, a thorough evaluation analysis was performed.

Data Curation

The data preparation was completed while considering the requirements of INCM. One requisite from INCM was that the model should not perform text simplification, it **should only extract the answer as it is written in the law**. This concept of extracting only the snippet of a document to answer the question is called Extractive Question Answering (Extractive QA). Another requisite from INCM was that the solution **should always return not only the law text, but also its metadata** (the article number, link to the legislation in dre.pt, related topics and references).

The goal of the Data Curation is to have a prepared input dataset where the model can extract its answers' from. As such, in the case of this project, the whole Portuguese Retirement Law and its metadata serves as the input dataset.

Before the group took over the project, the NOVA SBE Data Science Knowledge Center (DSKC) had already established contact with the INCM and had already received what was expected to be the full relevant law text: the complete consolidated Portuguese Retirement Law. While the text was originally presented as an XML file, the DSKC converted it into PowerPoint slides containing one law alinea and its metadata per slide. Additionally, a dataset of a series of historical user search inputs in DRE was also provided for the project. Later, INCM agreed that the user's comprehension of some specific Retirement Law articles required the understanding of specific non-Retirement Law articles, denominated External Statutes (ES). These articles, as well as their correspondence to Retirement Law articles, were also integrated into the dataset.

A Python pdf-mining library called PdfPlumber was used to transform the PowerPoint slides with the law text and all its metadata into a Pandas DataFrame. All the data was then compiled and stored in a single csv file which would later be used as the model's input. Each row of the csv file contained one article of the retirement law and its respective metadata, so that it could be retrieved and presented to the user with all this information.

Exploratory Data Analysis

INCM provided a file with more than a million queries done to its search engine (for all Law Statutes, not only for Retirement Law). This information was important because it allowed for the exploration of which and what types of queries were done by its users on the dre.pt. This EDA was relevant because these were the types of queries the solution was going to affect and whose search experience it should benefit.

The queries were filtered to contain the substrings “reform”, “pens”, or “aposent”, since these were, in principle, the queries that contained the keywords concerning Retirement Law. The queries related to retirement were analyzed and ranked by their frequency. This study can be found in Annex 2.

Two profiles of users were created: the ones that search specific articles (Law professionals) and the ones that perform broad queries (non-Law professionals). This study showed that most of the queries (77,1%) were deemed broad (see Annex 3), presumably having been made by a non-professional of law, which reiterates the importance of having a solution that is accessible to non-law professionals.

Further descriptive statistics were created to get a better understanding of the dataset. For more information on this analysis, see Annex 4-9.

A Knowledge Graph (KG) is a way to represent a network/collection of data points and illustrate the relationship between them in a more visual way. As such, a KG was created for a better understanding of the retirement laws and their relationships with each other. A Knowledge Graph is composed of nodes, labels, properties, and relationships. Nodes are data records that exist in the database: in this case, for example, a node is the topic called “Inscrição” that is labeled as a Topic, whose properties are the Topic_id and the Title of the respective topic. The topic has a relationship with multiple alineas. The devised KG can be seen in Figure 2.

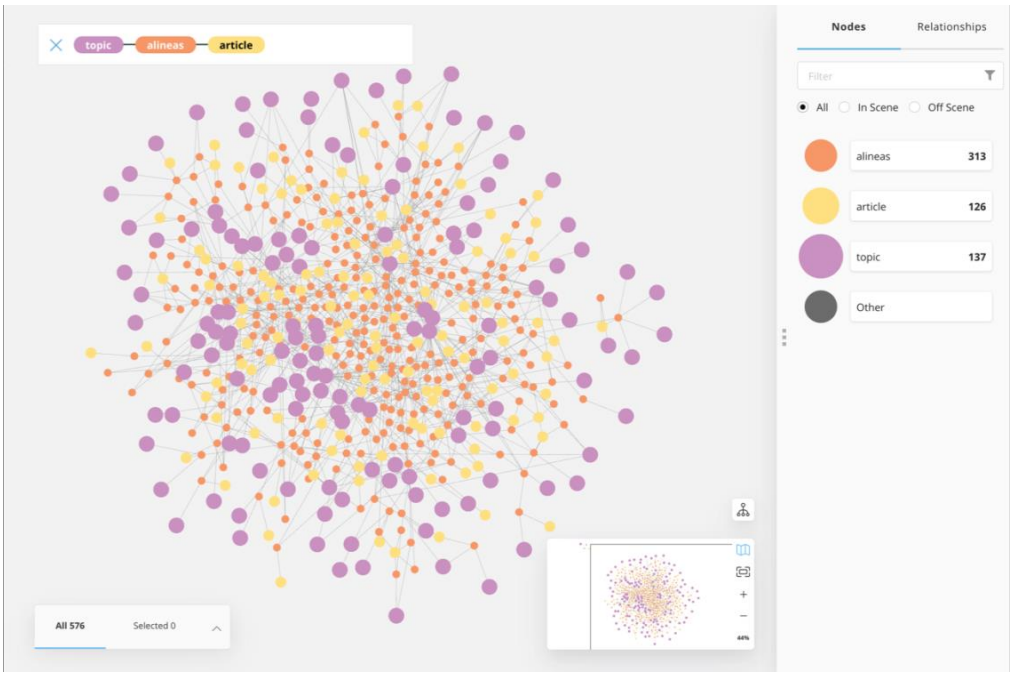


Figure 2 – Entire knowledge graph and relationships (right tab: number of nodes for each label)

The graph was built using the Neo4J software. The software allowed for the KG to be dynamic, allowed the data to be queried and transformed through its language Cypher and it also allowed the data and relationships to be exported in tabular form, making it ready to be used for a model.

The Knowledge Graph was important to get acquainted with the relationship between the documents that would later be fed to the model and to better visualize the data, allowing for a better understanding of the data. In Annex 10 and Annex 11, all nodes and their respective relationships are displayed. For more information on the KG developed, see Annex 12-17.

Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) is the state-of-the-art technique for various Natural Language Processing (NLP) tasks. Unlike previous NLP models that needed labeled data and thus making it hard to collect large quantities of data to train the models, BERT is a model pre-trained on unlabeled data, specifically thousands of millions of words from

BookCorpus (a large dataset made up of more than 11 thousand free novels) and Wikipedia. The original paper (Chang 2018) was developed and published by Google Brain. Google uses it in almost all of its NLP tasks, including search and translation. An example of an input question, input dataset and output of BERT is shown in Annex 18.

BERT was a great advance in various NLP tasks, mainly because of its bidirectionality feature. As opposed to other types of models that read the text sequentially (from left-to-right in most western languages, for example), BERT is Bidirectional, meaning that it takes into account the entire sequence of words, both from the left and the right, giving a much richer context to each word in the dataset because it takes into account all of its surroundings (see Annex 19 & 20 in the appendix for a better understanding of the difference between unidirectionality and bidirectionality). Through this feature, the model partially solves the problem of ambiguity, which had been one of the biggest limitations of previous models. (Kashnitsky 2019). Large pre-trained language models that can be applied to different language tasks are recent (BERT - 2019, GPT-3 - 2020), so they have some limitations. Explainability of the model is one of them, for example.

BERT has state-of-the-art performance in semantic search tasks. In semantic search the solution searches not only by keyword, but by intent and context of the question. Refer to Figure 1 in the Introduction Section for an example of this method.

Allowing users to search the Retirement Law in non-technical terms is exactly the goal of this project, so BERT was chosen for this project because it achieves state-of-the-art performance in semantic search and Extractive QA, it is simple to implement and easy to finetune. Also, unlike most Machine Learning models, there was no need for the creation of a training dataset in the beginning since BERT is already pre-trained in the Portuguese Language

BERT was the model chosen to be the pillar of this solution. More specifically, a version of BERTimbau (the Brazilian-Portuguese BERT) that is already fine-tuned for Question Answering tasks and achieves state-of-the-art performance (Guillou 2021). The model was obtained via the HuggingFace community (which allows to build, share, deploy and download open-source models, making BERTimbau available for download).

Haystack

A Question Answering system is often structured as presented in Figure 3. The solution takes a question from the user, searches for the right answer in its database and returns the document that answers the user's query.

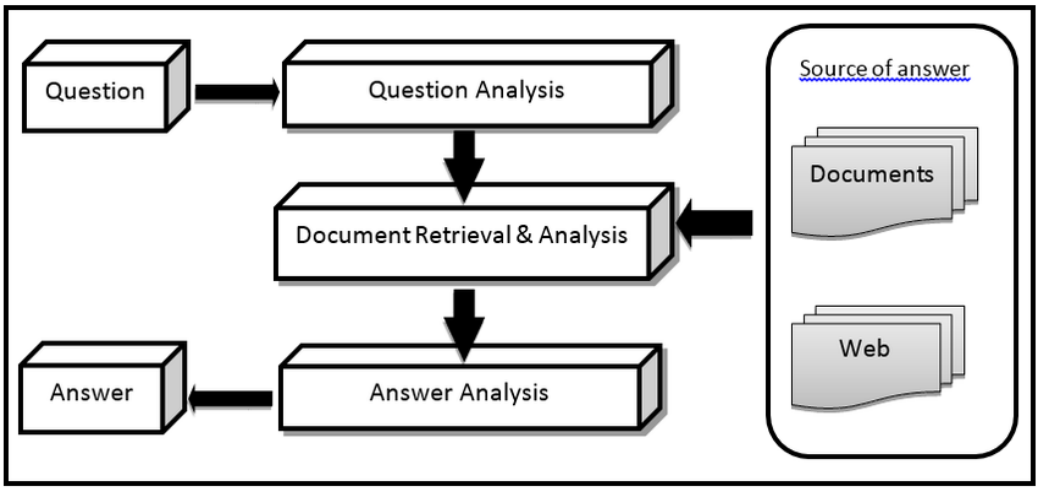


Figure 3 - Typical Architecture of a Question Answering System

Source: https://www.researchgate.net/figure/General-architecture-of-a-Question-Answering-System_fig1_292971999 accessed in December 2021

Haystack is an open-source framework that helps build end-to-end Question Answering systems using Deep Learning techniques to solve several Natural Language Processing tasks. Haystack's system is shown in Figure 4.

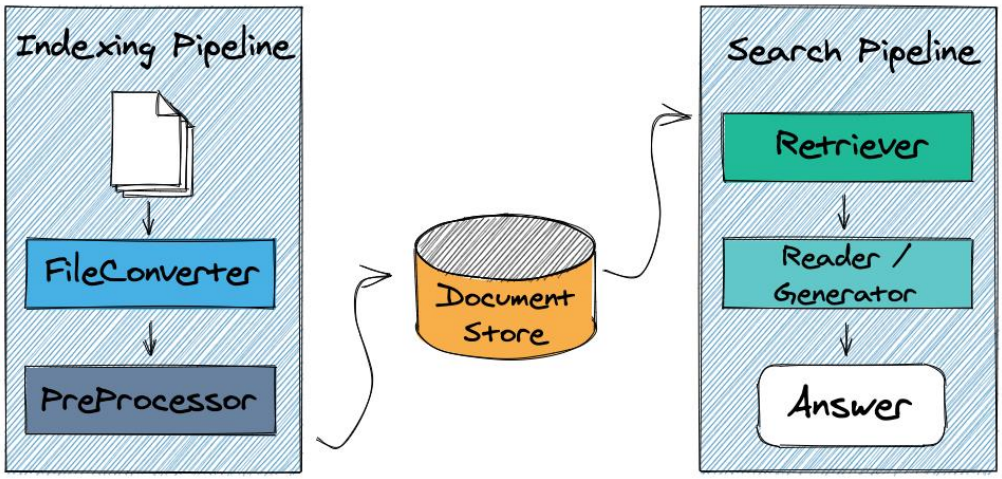


Figure 4 – Haystack Structure and Diagram of its pipelines

Source: https://github.com/deepset-ai/haystack accessed in December 2021

In the case of this project, Haystack's Indexing Pipeline pre-processes and stores the retirement laws, while the Search Pipeline processes the question from the user and retrieves an answer within the documents stored. As shown in Figure 4, Haystack has several stages in its pipelines.

The Indexing Pipeline in Figure 4 focuses on preparing (extracting, cleaning, splitting) the input dataset. The resulting text corpus is then stored in the DocumentStore (a database where the files are saved).

After the documents (laws) are stored in the DocumentStore, the system takes a query from the user. After this, the stored documents go through the evaluation of a Retriever.

The Retriever is a lightweight filter used to go through the whole database to pass on a list of documents that are answer candidates. This saves the Reader (which is much more heavy and needy in resources) from doing more work, speeding up the whole process as a result. After the list of most probable documents to answer the query is chosen by the Retriever, that list passes through the Reader.

The Reader is the core component that enables Haystack to find the right answer because it defines which answers should be returned to the user. The Reader can be any model that is already fine-tuned for Question Answering.

Haystack is an innovative tool in the field of NLP and is one of the more complete open-source frameworks in the field of Question Answering. It provides all the necessary methods to build an out-of-the-box semantic search solution: **easily customizable** components; **connected to the HuggingFace community** (allowing the use of BERTimbau as a Reader); possesses the most **recent language-agnostic Retrievers**; **minimal preprocessing** was needed to feed the input dataset into the model; developed with **scalability** in mind, so the input dataset can be composed of millions of documents; finally, it possesses an **annotation tool that allows the easy training of the model**. The tool also benefits from **quick support from its developers**, **good documentation**, **regular maintenance** and upgrades. Thus, Haystack was used to build the solution for INCM.

In the case of the present project, the documents stored and retrieved are the Retirement Laws. When a query is launched, the Retriever searches through all the documents and chooses a group of them (the recommended size for that group of documents by Haystack is 10). The Reader, BERTimbau, then analyses more in-depth each retrieved article and returns the specific part of the text that answers the question (usually just two or three phrases). Since INCM wanted the user to get more context, and for the model to retrieve the full article in which the answer was contained, the solution was modified to retrieve the full article text and all its metadata.

Performance Analysis & Evaluation of Extractive QA of Portuguese Retirement Laws

Context and Methodology

While most Extractive QA solutions are evaluated by their capability of returning specific expressions or phrases, INCM requested the solution to extract the full text of the correct articles instead. The implications of this goal modification are the following: since some queries are only answered by multiple articles, the system should return multiple articles if needed; it should also refrain from returning articles that do not answer the query, when possible, to allow a more frictionless search experience for INCM's users. Thus, it is important to evaluate the difference in performance when the model returns a different number of articles to the user. This will be the focus of this performance analysis.

The method used to evaluate the model was to compare the predicted articles of the solution with a testing set composed of 180 questions and answers about Retirement Law through several evaluation metrics used for information retrieval.

Evaluation metrics of information retrieval systems measure how likely each group of documents is to correctly answer the user. The evaluation metrics were used to measure the difference in performance of the solution when it returns different amounts of articles. They will also allow INCM to quantitatively compare the performance of future iterations of the system with this baseline solution.

The following metrics were used since they are a standard for information retrieval solutions and are the most pervasive ones in the context of information retrieval systems (Manning 2009):

1. ***Exact Match***: percentage of how many times the correct articles are retrieved as one of the answers.

2. ***Recal@K***: the probability that a relevant document is retrieved by the query (*successful_documents_returned / successful_documents_supposed_to_return*);
3. ***Precision@K***: fraction of the documents retrieved that are relevant to the user's information need (*number_of_right_topics_predicted / number_of_topics_predicted*);
4. ***F1-Score***: Harmonic Mean of the Recall@K and Precision@K;
5. ***Mean Reciprocal Rank***: for each question, the model gets a score based on answers' order (ex: 3 answers were returned. If the right article is the first article retrieved, then the score is 1. If the right article is only the third returned, the MRR is ⅓).

Both the Retriever and the Reader were first evaluated separately. On one hand, the different Retrievers were compared to each other and the one that achieved best performance in the tests was chosen to be the Retriever of the solution. The only metric used to evaluate the Retriever was Exact Match. On the other hand, since the only Reader available in Portuguese is BERTimbau, BERTimbau's performance was evaluated when it returned 1 to 10 articles to the user. Every metric previously mentioned was used to evaluate the Reader. Finally, an overall evaluation of the solution was performed.

Retriever Choice and Analysis

As mentioned in the "Haystack" Section, the Retriever is a filter that has one goal: it should be as lightweight as possible, and it should go through all the documents and point out those that are relevant to the query. The Retriever serves the purpose of excluding the obvious negative cases (cases where the documents obviously don't answer the query given), such that the Reader has to be used only a limited number of times, thus improving the speed of the Question Answering process.

Retrievers can be put into two main categories: Sparse and Dense. Sparse Retrievers look for shared keywords between the document and question. They are simple to implement, computationally cheaper than their Dense counterparts, don't need to be trained and they work on any language. They are effective and have very good performance (Turnbull, 2015).

In turn, Dense solutions show even better performance than Sparse, but are also more computationally expensive, especially during indexing (Athar & Ali 2021, Facebook AI 2019). Dense retrievers are trained on labeled data, and they are also language specific. While the English DPR can still work well with other languages in some cases, it does not reach the same level of performance achieved in English. Nonetheless, since there is no Portuguese DPR available, in this analysis the English DPR was used to compare to the Sparse alternatives (which are language agnostic).

The Haystack framework provides three Retrievers to choose from: Term Frequency–Inverse Document Frequency (TF-IDF), Best Matching25 (BM25) and Dense Passage Retrieval (DPR).

The results of the tests performed are shown in Figure 5. DPR achieves 0 correct answers, TF-IDF achieves 140 correct answers and the BM25 achieves 143 correct answers.

Figure 5 shows that the Sparse Retrievers have a much better performance (showing an average of 141.5 correct answers), than the DPR (which achieved 0 correct answers). This was expected since Sparse Retrievers are language agnostic. Sparse solutions are better suited for getting very good results with no resources allocated to training, allowing for more time and resources to be spent elsewhere. Also unsurprisingly, when comparing the two sparse Retrievers, the BM25 had a better performance than TF-IDF. This was expected since BM25 is the

recommended Retriever choice by Haystack, shows state-of-the-art results and it is built upon TF-IDF (Seitz 2020).

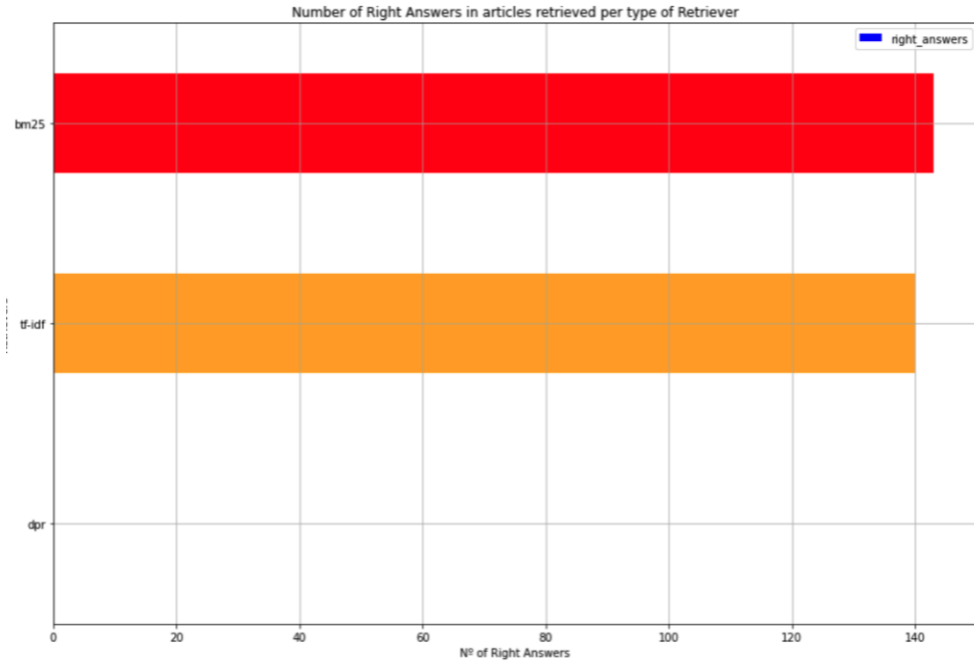


Figure 5 - Comparison between the performance of the three (DPR - Blue, BM25 - Red, TF-IDF - Orange) Retrievers available in Haystack

As such, the BM25 was the chosen Retriever for this project because of its results in the tests, its speed of performance and it effectively represents the state-of-the-art retrieval functions used in document retrieval.

Reader Choice and Results' Analysis

As discussed in the "Haystack" Section, the most important part of this QA system is the Reader since it enables Haystack to find the right answers between all the documents returned by the Retriever.

As previously mentioned in the "BERT" Section, the Portuguese BERT was chosen to be the Reader of this project since it achieves state-of-the-art results on QA tasks (Souza & Nogueira

& Lotufo 2021) and is the only deep bidirectional transformer model that is freely available and already pre-trained in Portuguese.

After choosing BERTimbau as the Reader, the performance analysis of the model was conducted. As it was previously mentioned in this Section, all the mentioned evaluation metrics were used to analyse the performance of the model. Table 1 shows the results of the performance of the model when the Reader returns from 1 to 10 documents per question.

Table 1 - Performance of BERTimbau with different number of articles retrieved on the testing dataset.

top_k_reader	Exact Match	Recall	Precision	MRR	F1-Score
1	22%	0.229391	0.215054	0.236559	0.221991
2	33%	0.349910	0.233871	0.319892	0.280358
3	41%	0.435484	0.220430	0.380824	0.292702
4	48%	0.514337	0.206989	0.414875	0.295185
5	55%	0.582885	0.213978	0.439247	0.313039
6	62%	0.649194	0.212366	0.461111	0.320039
7	68%	0.709229	0.211982	0.474859	0.326404
8	70%	0.728047	0.202285	0.480556	0.316603
9	72%	0.743280	0.200119	0.485633	0.315338
10	73%	0.762097	0.196237	0.493235	0.312107

As presented in Table 1, the more articles returned by the Reader, the more correct articles are returned. When the number of returned articles is 2, for example, 33% of the queries had all the right articles returned. When the number of articles increases to 55,5% of the queries were fully answered by the articles returned to the user. The Exact Match, Recall@K and MRR increase every time the number of retrieved documents increases.

Precision, however, gets worse the more articles are returned. This behavior is expected because most questions in the testing set only have one topic associated with them. By definition,

when the right answer has only one topic, but the number of topics returned increases, the precision will decrease. Take the example of the Reader returning 2 documents that have different topics associated with them. If the question was only one related to one of the two topics, Precision would be ½. If the same question had 5 articles returned with 5 different topics associated to them, Precision would be ⅕.

The F1-Score peaks when the Reader returns 7, since the metric is the harmonic mean of the Recall@K and Precision@K. In other words, it considers the number of right documents retrieved, but at the same time penalizes if the solution returns too many unrelated articles to the user. For this reason, the F1-Score is often used as a summary of the evaluation metrics. Thus, it was considered one of the most important metrics for the choice of the Reader.

Another important metric is the incremental increase of Exact Match, presented in Figure 6.

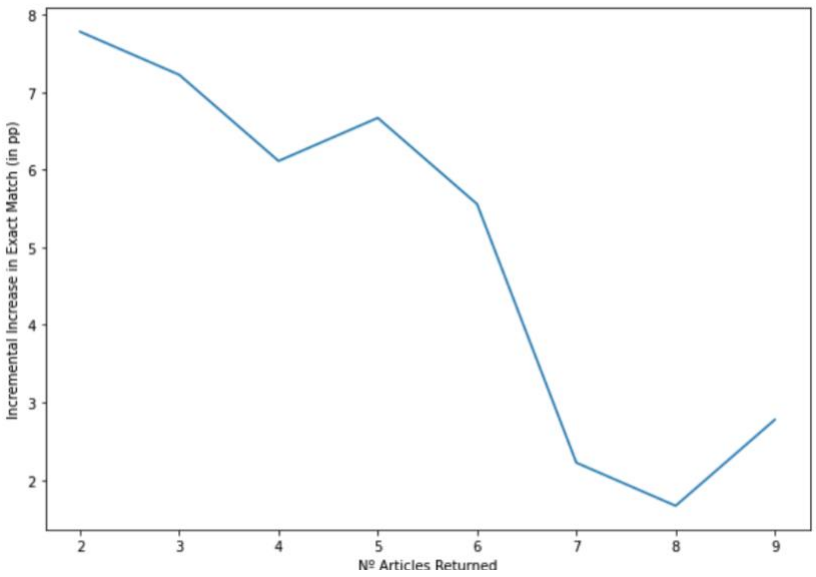


Figure 6 - Exact Match Incremental Increase with different number of articles retrieved

The metric starts to decrease significantly after the Reader returns 5 articles and eventually stagnates after 7 articles. This means that, from 7 articles returned upwards, for each increase of

articles returned per question, each increase would give no more than a 2% increase in questions fully answered.

When returning 5 to 7 articles per question, the model achieves the highest F1-Score and it is also where the incremental increase of the Exact Match starts to decrease and eventually stagnate.

Testing Platform

A prototype web-app was created with React and FastAPI to present the solution and to allow INCM to perform tests on the model.

The app created was a simple web page that allowed INCM's jurists to perform a variety of actions: input a question about Retirement Law and get an answer back, read the already existing laws in a user-friendly way and read the previous queries, the previous answers given and the respective feedback the answer got from the user. The backend was managed by an API created with FastAPI, which allowed the model to run easily and quickly.

This testing platform was essential for INCM's team of jurists to evaluate the potential of this solution and its limitations. It was also relevant for the project to build various iterations of the solution and see what works best from a product point-of-view.

The solution was well received by the group of jurists, with all the tests receiving overall positive feedback by the legal team of INCM.

At the end of the tests, a form was distributed for the two members of INCM's legal team that were responsible for testing the platform. The form asked them to evaluate the performance of the prototype solution and state their own expectations for the implemented technology. Both said the solution had potential to be used in the future and that the solution would be beneficial for more Law domains. Both indicated that the solution performed better than INCM's current

solution (providing a rating of 4 and 5 to the solution). When asked how useful this solution could be for the non-Law professional, the jurists also provided a rating of 4 and 5.

This was important because it validates what the performance analysis had discovered: the model is showing promising results and can, in fact, help the overall Portuguese population get more informed about the Retirement Law.

Comparison with INCM's Current Solution

INCM's current search tool has a variety of limitations: it searches by keyword instead of taking the context and intent of the user into account; it does not account for orthographic errors; and it also does not perform semantic search (queries similar to "When can I retire?" and "What is my income when I retire?" do not get any response from the site).

The Portuguese law has a very specific type of language that a non-Law professional often does not use, so the current solution implemented is inappropriate to the overall population's needs.

In contrast, the solution that was developed for this project allows for: semantic search, considers (and works with) orthographic errors and functionally understands synonyms. Thus, the solution developed in this study solved the most important limitations of the current solution.

In the 'Testing Platform' subsection, the form distributed to the legal team of INCM highlights that the prototype solution created in this project already shows potential and already performs better than their current solution.

Discussion and Overall Performance of the Solution

The performance analysis of the Question Answering pipeline showed that the Portuguese version of BERT allied with the Haystack Framework has potential for complex Extractive QA

tasks. It presents positive results right out-of-the-box, i.e., without any further training besides the initial pre-training in Brazilian Portuguese.

When retrieving only 7 articles per question, the model returned all the articles necessary to correctly answer the question 68% of the time, which was also considered an advance when compared to the current solution by INCM's legal team.

However, as expected, when comparing these results with the original results of BERTimbau, this solution showcases a worse performance, since it was not trained on Portuguese Retirement Law. It is important to note that when it was trained with only 87 questions and answers related to Retirement Law, the model already achieved better results than its baseline.

Limitations

One limitation of this prototype is the lack of training of BERTimbau on Portuguese Law. The model performs worse in this case study than with the Brazilian Portuguese dataset it was initially tested against because the Reader was not trained in Portuguese law-specific data. The lack of time to perform this training did not allow the trained model to be implemented for these tests.

Comparing the results of the solution with other Extractive QA systems is also one of the limitations of this performance analysis. The comparison with other similar systems is difficult because this solution must retrieve the full article(s) to be correct, while other solutions are evaluated on having to only return a phrase or expression that correctly answers the question. Furthermore, other studies available on the domain of Portuguese QA are Open-Domain QA systems, and not Closed-Domain like this one, making it difficult to make a fair comparison.

Another limitation of the solution is its explainability, more specifically the explainability of BERT. However, it is important to mention that the 'black-box' problem with large Deep Learning models is not exclusive to this solution.

The size of the testing set not being big enough is also one of the limitations of this analysis. Ideally, the model would be tested against thousands of questions about retirement law, not only 180 questions.

The number of answers returned being a fixed number is another limitation of this performance analysis. Ideally, the model would return only the needed articles to answer a certain query. For example, if the query only needs two articles to be returned, it should return only those two articles, not a fixed number of them for every query. This also makes the overall results of the model decrease, especially the metric Precision@K and F1-Score.

Finally, this version of the model still does not have a full feedback feature implemented. This tool would allow for the model to be trained through the users' feedback, making it perform better the more it is used.

Conclusion

This research project aimed to analyze if BERT could be a reliable solution for improving the search task of the Portuguese Retirement Law. More specifically, this thesis aimed to analyze and evaluate if BERT could be a potential solution for an Extractive QA/semantic search task that would be accessible for all the Portuguese population.

Based on a quantitative analysis of its performance on the testing set provided by INCM, it can be concluded that BERT has potential. With only a prototype, it showed positive results as a search tool for the Portuguese Retirement Law (68% Exact Match when retrieving 7 articles).

Furthermore, the jurists' team of INCM stated this prototype solution is more accessible for the overall Portuguese population when compared to the current search solution in dre.pt. Moreover, it surpassed many of the limitations of INCM's current solution, the main ones being that it can be used by non-Law professionals, it allows for semantic search instead of keyword search and allows for orthographic errors. This project developed a solution that can serve as the initial step for INCM's next search solution.

Since the proposed solution is still in a prototype, it has some limitations, the biggest one being that it is not trained specifically in Portuguese Law, which negatively affects its performance.

To conclude, the overall results indicate that BERT can in fact be a powerful out-of-the-box tool for Extractive Question Answering in relation to Portuguese Retirement Law, solving the limitations of the current solution of INCM.

Future Work

The biggest limitation of the model is not being trained on law-specific data. In the future, the model should be trained with questions and answers related to the retirement law. Some preliminary tests show that a training dataset composed of 87 questions and answers would already make the model achieve better results. This training can be performed through the creation of a similar dataset to the one that was given for testing. Haystack's framework provides a simple way to perform this task, either through its more manual annotation tool or through an automatic user feedback restAPI.

Another feature that can be explored in the future is the increase of the scope of the project to include the rest of the Portuguese Law, not only the Retirement domain. In a way, if there is

enough time to train and fine-tune the model, there is no reason as to why the scope of the project should be restrained to only the Retirement Law. This way, the solution would prove even more powerful and more useful to the overall Portuguese population. This should be the end goal for this project, as it would make the Portuguese Law more accessible and democratize this knowledge and information.

In the same vein, it would be interesting to test this kind of solution in various fields besides law. This type of Extractive QA solution could be completely disruptive to various industries. The medical science, biology, history, physics, and pharmaceuticals industries, for example, could be changed completely if BERT can have a satisfying performance when used for QA in relation to complex and very technical reports. The customer service platforms of all industries are also already being affected by this type of solution.

References

[1] Chang, Ming-Wei, Jacob Devlin, Kenton Lee, and Kristina Toutanova. “Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” arxiv.org, October 11, 2019. https://aclanthology.org/N19-1423.pdf.

[2] Ionita, Matei, Yury Kashnitsky, Ken Krige, Vladimir Larin, Atanas Atanasov, and Dennis Logvinenko. “Resolving Gendered Ambiguous Pronouns with Bert.” Proceedings of the First Workshop on Gender Bias in Natural Language Processing, 2019. https://doi.org/10.18653/v1/w19-3817.

[3] Guillou, Pierre. “NLP: Modelo De Question Answering Em Qualquer Idioma Baseado No Bert Base“ Medium. Medium, June 18, 2021. https://medium.com/@pierre_guillou/nlp-modelo-de-question-answering-em-qualquer-idioma-baseado-no-bert-base-estudo-de-caso-em-12093d385e78.

[4] Manning, Christopher D., Hinrich Schütze, and Prabhakar Raghavan. “Online Edition (C)2009 Cambridge up - Stanford University.” stanford.edu. Cambridge University Press, 2009. https://nlp.stanford.edu/IR-book/pdf/08eval.pdf.

[5] Turnbull, Doug. “BM25 The next Generation of Lucene Relevance.” OpenSource Connections, September 9, 2020. https://opensourceconnections.com/blog/2015/10/16/bm25-the-next-generation-of-lucene-relevation/.

[6] Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. “Dense Passage Retrieval for Open-Domain Question Answering.” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. https://doi.org/10.18653/v1/2020.emnlp-main.550.

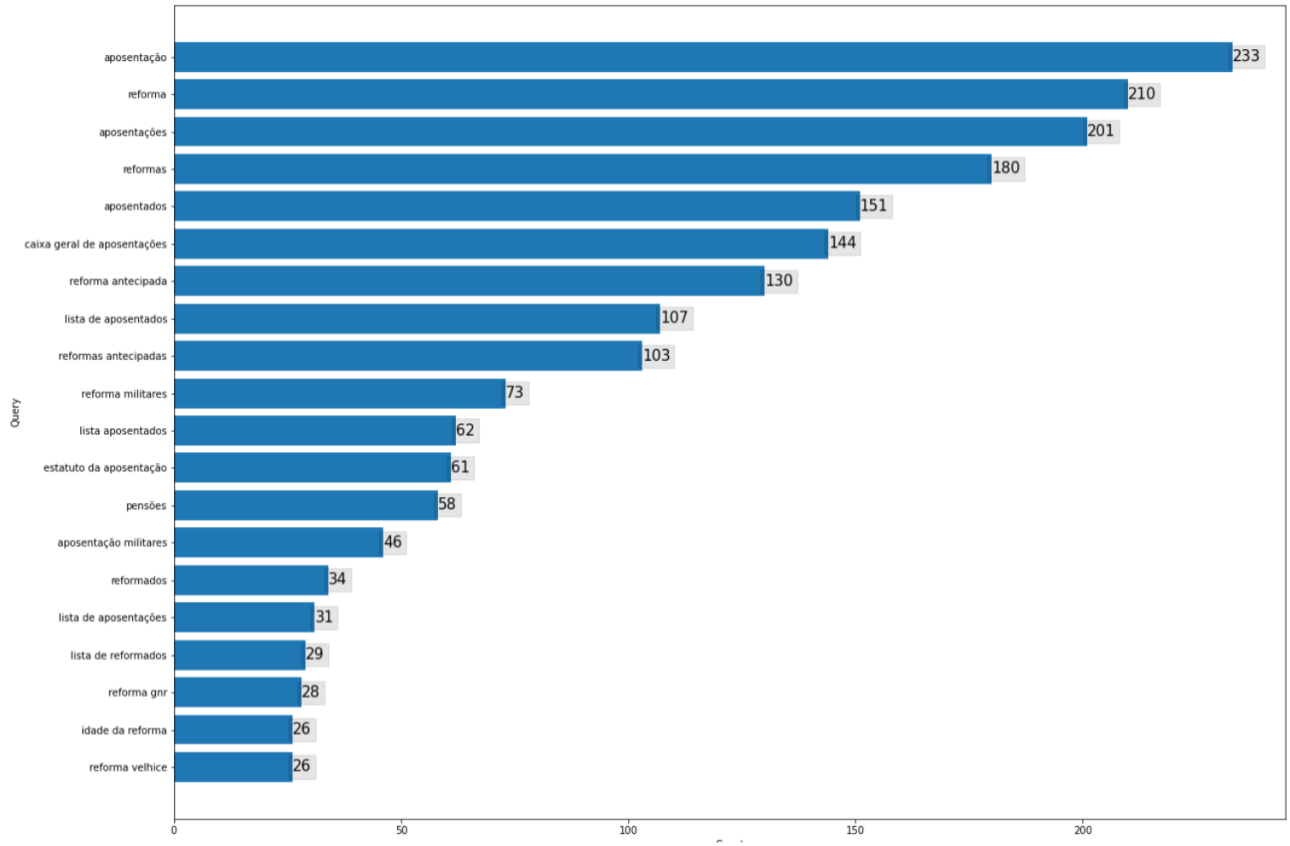
[7] Seitz, Rudi. “Understanding TF-IDF and BM-25.” KMW Technology, March 20, 2020. https://kmwllc.com/index.php/2020/03/20/understanding-tf-idf-and-bm-25/.

[8] Souza, Fábio, Rodrigo Nogueira, and Roberto Lotufo. “Bertimbau: Pretrained Bert Models for Brazilian Portuguese.” Intelligent Systems, 2020, 403–17. https://doi.org/10.1007/978-3-030-61377-8_28.

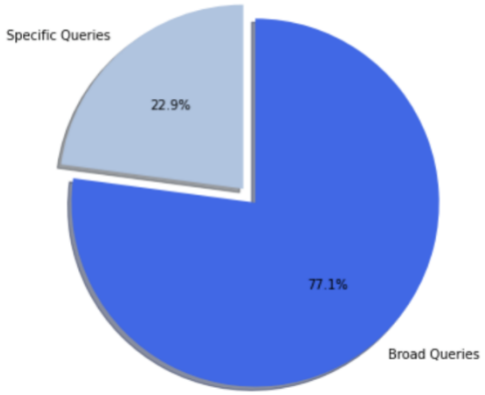
Appendix

Document	Attribute name	Interpretation
2020_10 INCM_Annotation_consolidado	Title	This attribute identifies the title of the article on which the annotation is about
2020_10 INCM_Annotation_consolidado	Article Number	This attribute identifies the article on which the annotation is about, using its number
2020_10 INCM_Annotation_consolidado	Alinea	This attribute identifies the alinea of the article to which the law text refers
2020_10 INCM_Annotation_consolidado	Revisto por:	This attribute defines who reviewed the annotation
2020_10 INCM_Annotation_consolidado	Validado:	This attribute defines who validated the annotation
2020_10 INCM_Annotation_consolidado	Law Description	This attribute is the article as it is written in the Law
2020_10 INCM_Annotation_consolidado	Tópico:	This attribute assigns different Law articles to a given topic
2020_10 INCM_Annotation_consolidado	Referências:	This attribute identifies references to articles from *Estatuto da Aposentação*
2020_10 INCM_Annotation_consolidado	Ligações	This attribute relates each article to broader areas/topics of the Law
2020_10 INCM_Annotation_consolidado	REFERÊNCIA INCM:	This attribute identifies references to articles outside of *Estatuto da Aposentação*

Annex 1 – Data Dictionary of Dataset of Retirement Laws provided by INCM



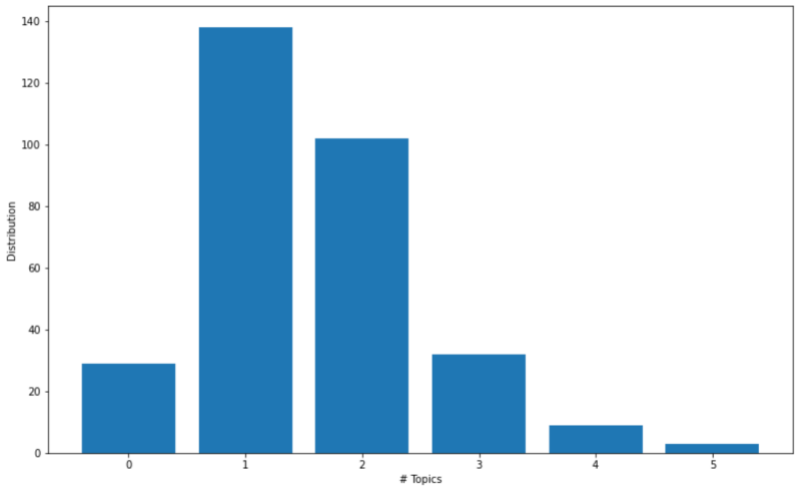
Annex 2 - Frequency Distribution of Queries (top 20)



Annex 3 - Percentage of Specific vs. Broad Queries on DRE.pt

Metric	Value
# Topics	125
Maximum # topics per article	5
Minimum # topics per article	0
Mean # topics per article	1.56
# Alineas without topics	29

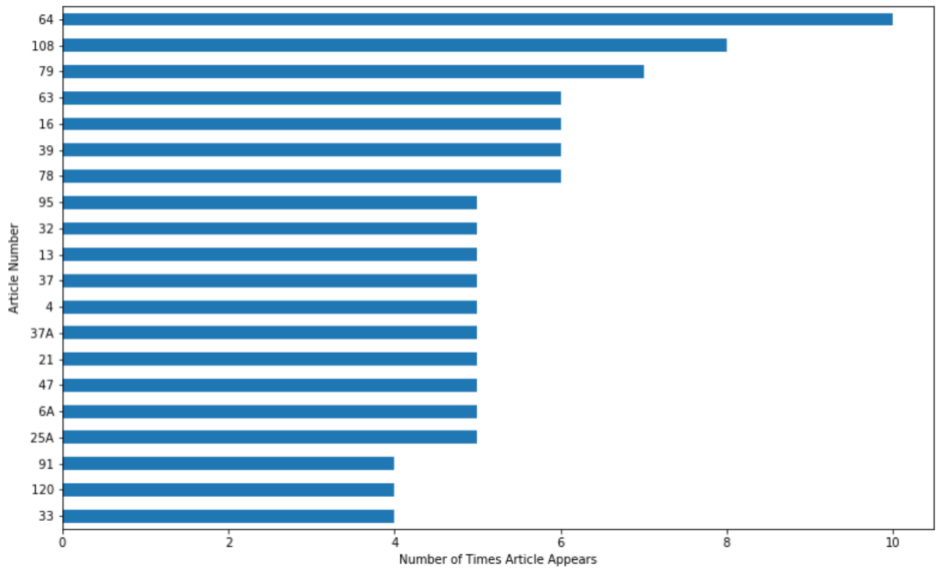
Annex 4 - Topic-related descriptive statistics



Annex 5 - Frequency distribution of topics per alinea

Metric	Value
# Articles	126
# Alineas	313
Mean # alineas per article	2.48
Mean # characters per alinea	251.86
Mean # words per alinea	40.76
# Revoked alineas	3

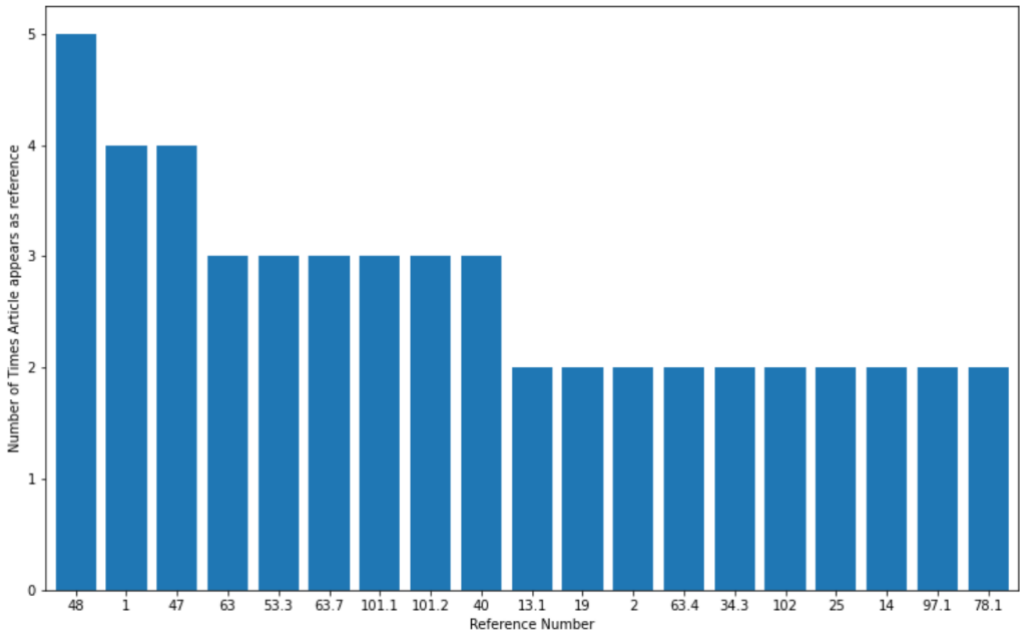
Annex 6 - Alinea-related descriptive statistics



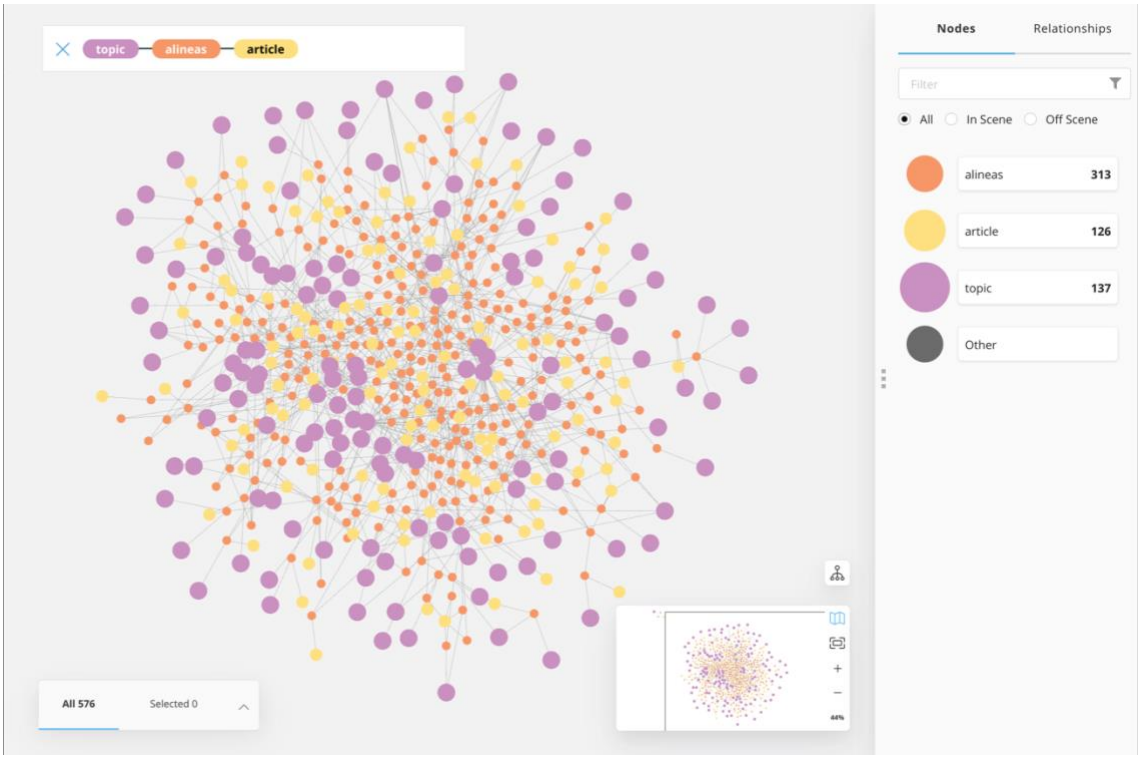
Annex 7 - Articles with the most alineas (top 20)

Metric	Value
# Retirement law references	252
# External law references	94
# Alineas without references	134
Mean # references per alinea	1

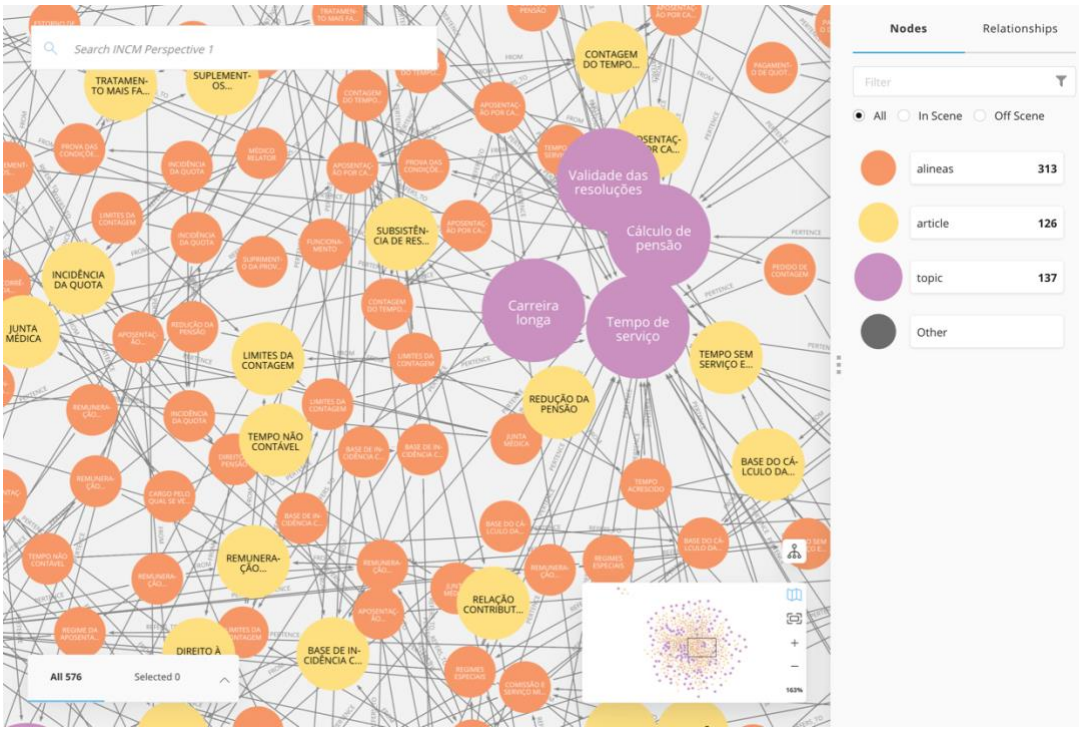
Annex 8 - Reference-related descriptive statistics



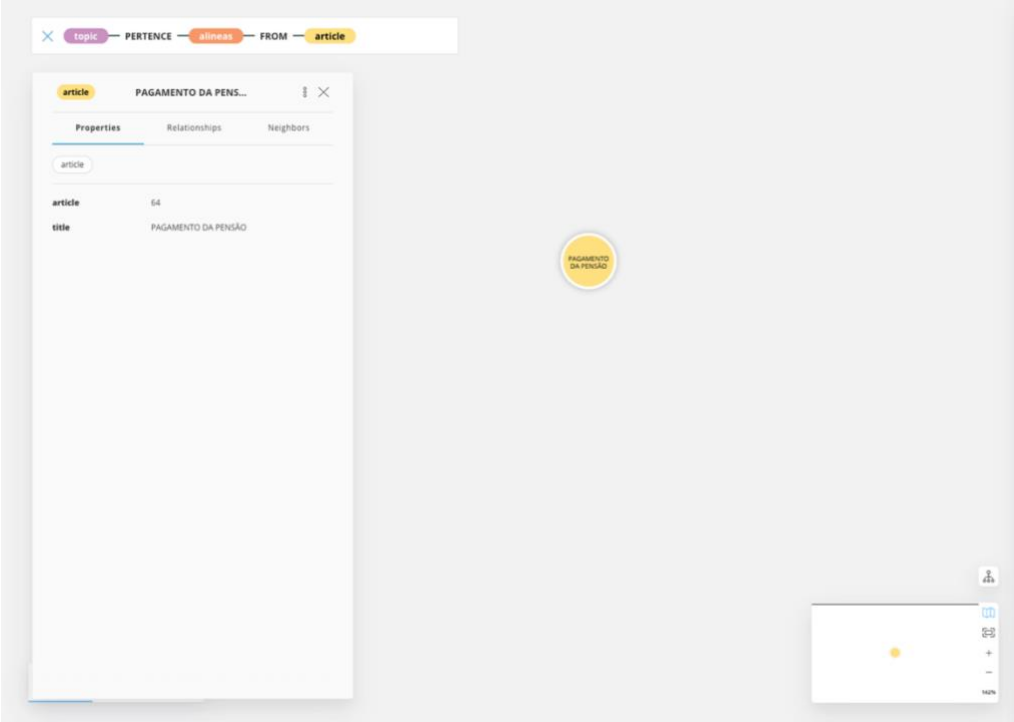
Annex 9 - Number of times the articles are referenced



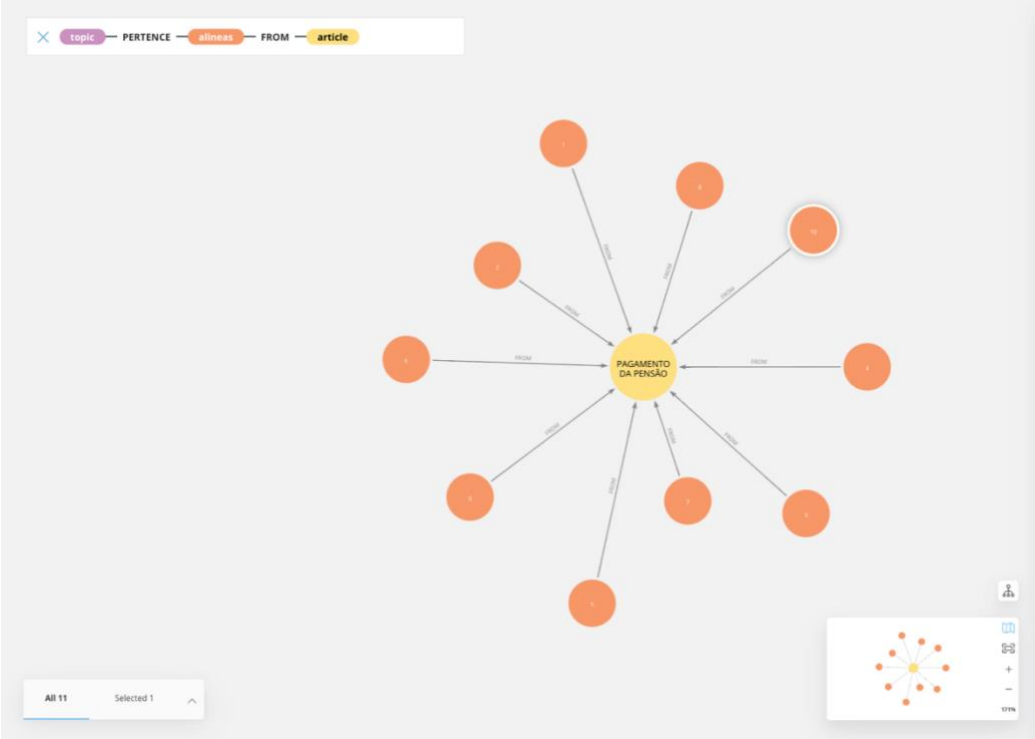
Annex 10 - Entire knowledge graph and relationships (right tab: number of nodes for each label)



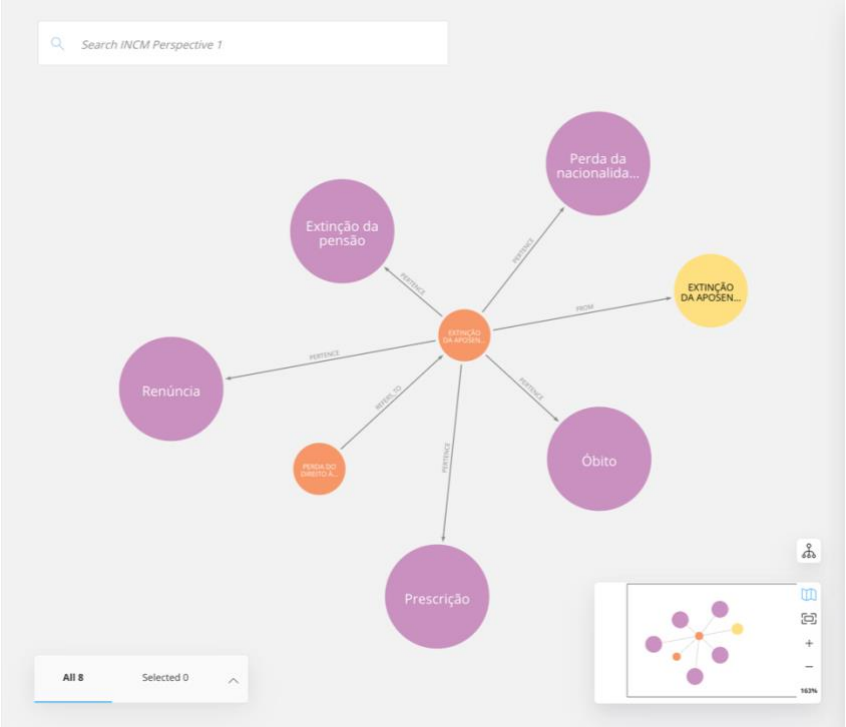
Annex 11 - Knowledge graph and relationships (zoomed in)



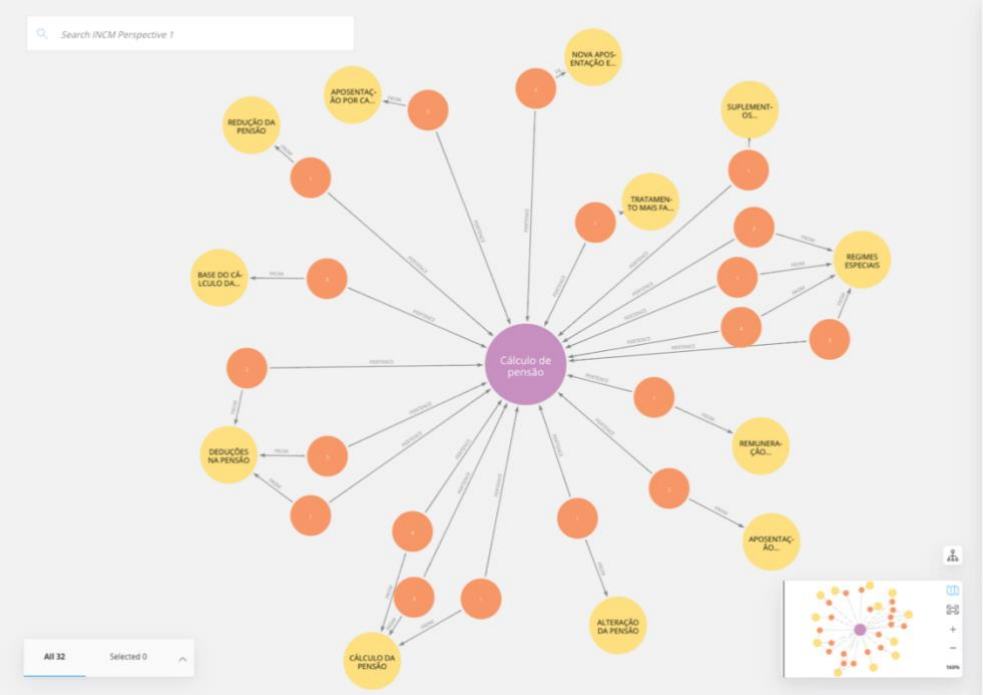
Annex 12 - Only the article "Pagamento da Pensão" and the respective properties (article and title)



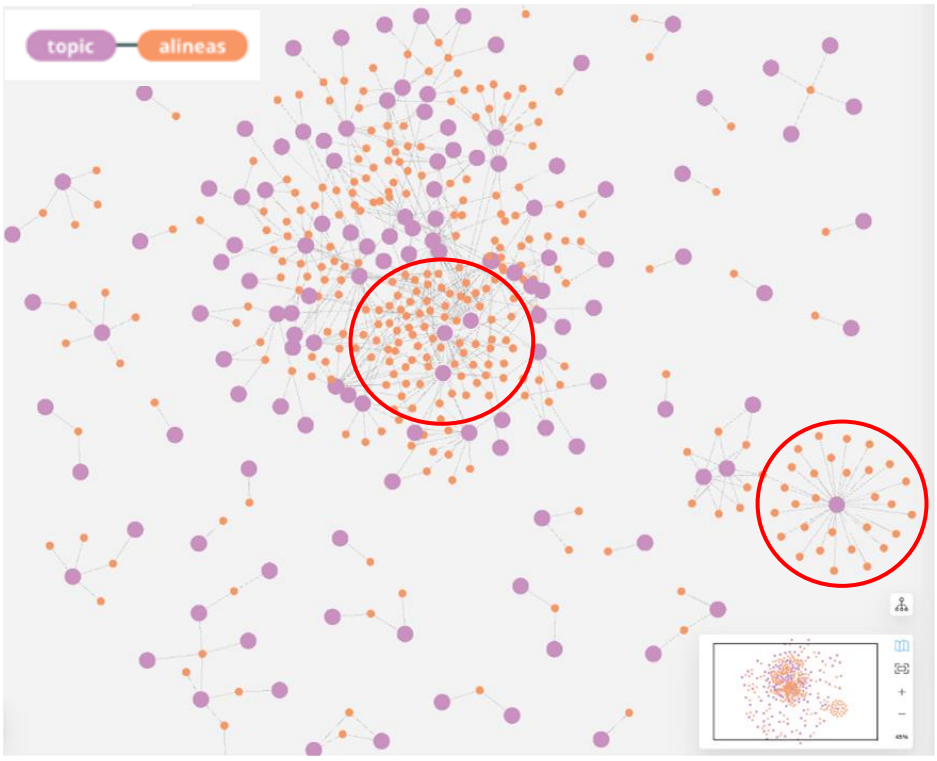
Annex 13 - All alineas of article "Pagamento da Pensão"



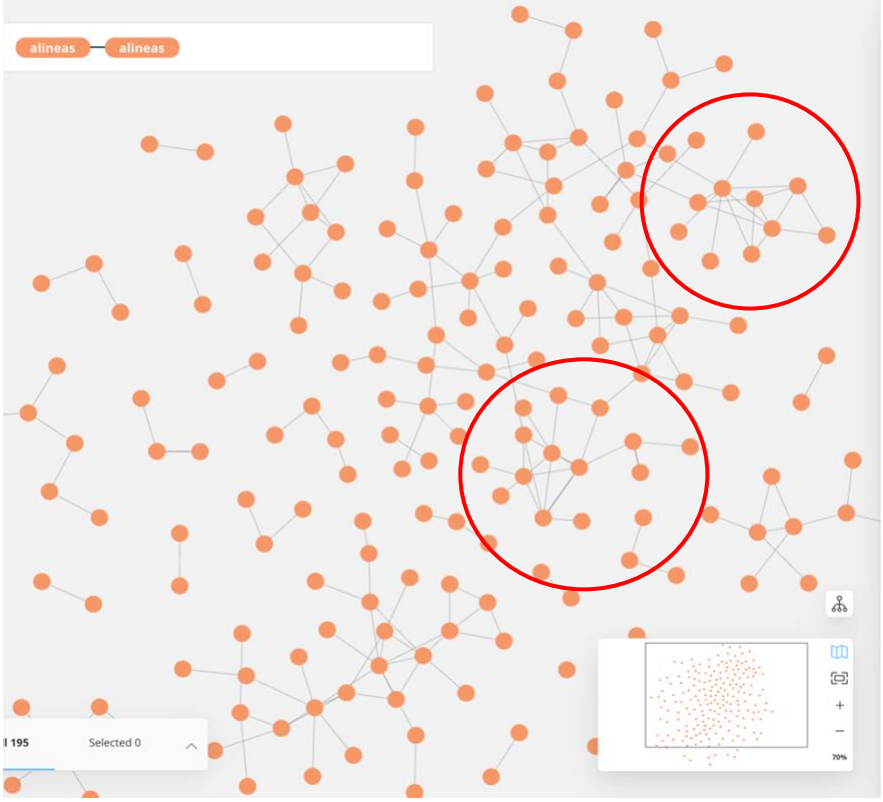
Annex 14 - All related nodes and relationships to alinea 1 of article 82



Annex 15 - All alineas related to to the topic 'Cálculo de Pensão' and all their mother-articles



Annex 16 - The 3 topics with the most alineas (left circle) and alineas without any topic (right circle)



Annex 17 - Clusters of alineas with the most references

- Input Question:

  Where do water droplets collide with ice crystals to form precipitation?

- Input Paragraph:

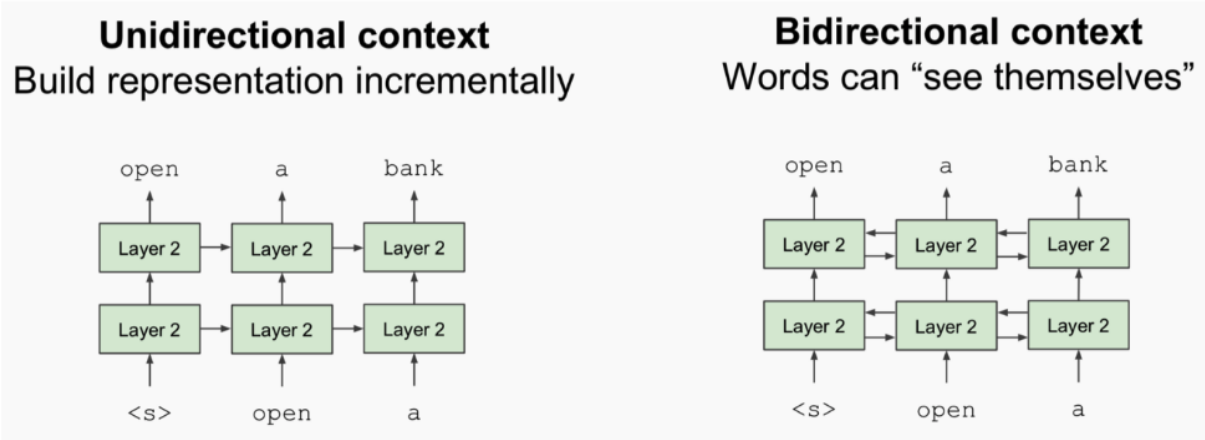
  ... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. ...

- Output Answer:

  within a cloud

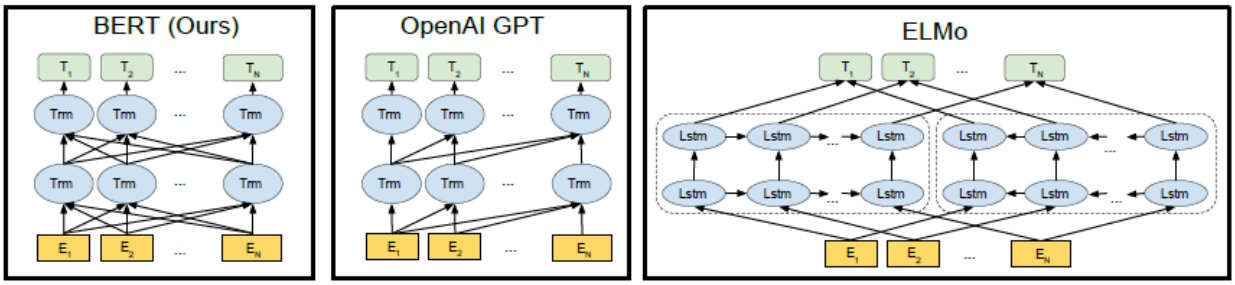
Annex 18 - Example of an input question, input dataset and output of BERT

Source: https://medium.com/@samia.khalid/bert-explained-a-complete-guide-with-theory-and-tutorial-3ac9ebc8fa7c accessed in December 2021



Annex 19 - Unidirectional Context vs Bidirectional Context

Src: https://www.cs.princeton.edu/courses/archive/fall19/cos484/lectures/lec16.pdf accessed on December 2021



Annex 20 - Bidirectionality advances of BERT compared with Unidirectionality of OpenAI's GPT and ELMo models

Source: https://medium.com/swlh/bert-bidirectional-encoder-representations-from-transformers-c1ba3ef5e2f4 accessed on December 2021