ALEXANDRE ALCOFORADO
RODRIGO ELIZARDO GERBER

ZEROBERTO GOES TO THE PARLIAMENT: ZERO-SHOT CLASSIFICATION APPLIED TO THE ASSEMBLY OF THE PORTUGUESE REPUBLIC

São Paulo
2021

ALEXANDRE ALCOFORADO
RODRIGO ELIZARDO GERBER

ZEROBERTO GOES TO THE PARLIAMENT: ZERO-SHOT CLASSIFICATION APPLIED TO THE ASSEMBLY OF THE PORTUGUESE REPUBLIC

Trabalho apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Engenheiro.

São Paulo
2021

ALEXANDRE ALCOFORADO
RODRIGO ELIZARDO GERBER

ZEROBERTO GOES TO THE PARLIAMENT: ZERO-SHOT CLASSIFICATION APPLIED TO THE ASSEMBLY OF THE PORTUGUESE REPUBLIC

Trabalho apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Engenheiro.

Orientador:

Anna Helena Reali Costa

São Paulo
2021

SUMÁRIO

21 de dezembro de 2021

1 INTRODUCTION AND MOTIVATION

The world has gone digital. The development of new technologies - and the speed by which they get widespread and are adopted by societies - has increased dramatically in the last 30 years [28]. This phenomenon has brought improvements to many aspects of citizens daily life, causing many common activities to turn digital: social networking, medical consultations, shopping and even asking for a cab.

There is evidence to believe this process is not over [29]. New technologies will continue to emerge, and their adoption by common people should keep its quickness. Within this context, many opportunities for engineering are presented: connectivity has been raised to a new level, with promising technologies such as autonomous cars and trucks on the horizon [10]. Social networks like Twitter, WhatsApp, Facebook and Instagram have hundreds of millions - or billions - of users worldwide, and many of these users are remotely connected all the time. This promotes extremely fast dynamics in these networks: in them, users receive an enormous amount of information per second, and it is their task to manage, filter and select this information. This causes a scenario of information overload [23, 3].

Although life in this digital era has streamlined in many ways, politics of most countries are still conducted in the same fashion as before: information about the democratic processes lack the power of synthesis present in most digital content. There is, thus, a mismatch between the citizens' digital lives and the political behaviour of their countries.

The difficulty of accessing public information is aggravated in the context of social networks: if on the one hand, they have brought many improvements to human life, on the other hand, they facilitate the spread of disinformation and conspiracy theories. Thus, malicious groups can generate social tension and polarization, disseminating fake news or extremist speeches, and may even influence election results [35].

Well-informed citizens and spaces for debate and criticism are fundamental principles of a mature democracy. Without them, the quality of public debate, that societies must

go through to improve themselves, may degrade [27]. With that in mind, it is clear that the problems presented are among the factors that significantly contribute to the current reduction of trust in democratic institutions [33].

So, the concept of **Digital Democracy** or e-democracy gains strength: it can be understood as the use of technology in the policy formulation process and citizen-state relations by creating tools that encourage direct citizen participation in the decisions and discussions societies must go through [7]. Digital democracy is defined in three axis: information, discussion and participation. Improving the quality of information is, then, addressing the first axis of the digital democracy. We argue there is an opportunity for applying new technologies to public data: by using them as a service for democracy, we can bring politics closer to the citizenry and promote debates on the most relevant issues of the present days.

It is noticed that the information present in these social networks is *noisy* in the sense that there is much more data available than any individual user can consume. However, information produced by state agencies, in general, is also *noisy*: it often consists of documentation that is extensive and difficult for citizens to interpret.

The term *big data* has had different definitions until it became widespread, around 2011. Since then, effort has been made in the direction of defining, conceptualizing and developing methodology for analyzing it. [11] offers a broader definition for the term, citing the 3 V's as main criteria: Volume, Velocity and Variety. Another important criterion proposed is the need for sophisticated, innovative techniques and technologies for processing it. In this work, we argue that data from state agencies is, thus, big data, since they are information assets with high-volume, high-velocity and high-variety, which require innovative techniques for capturing, storing and processing the information.

1.1 Low-Resource Natural Language Processing

In the last years, it is noteworthy that there has been a revolution in the Artificial Intelligence field. Recent technological – software and hardware – breakthroughs have brought discoveries of new tasks that machines can perform. In the Natural Language Processing (NLP) field, for instance, techniques already have the power to automatically interpret large volumes of information [39]. This kind of ability enables the extraction of relevant and straightforward information from a large number of documents, many of which may be composed of long and complex texts.

A lot of applications for NLP are present in citizens daily life. Applications like Google Translate, Grammarly and Alexa are very popular worldwide, but machines still fall behind in many tasks that humans can do. However, machines have an unbeatable advantage: they can perform tasks in large scale. Therefore, classifying thousands – or millions – of documents into classes or separating them into topics, for instance, even though it can be hard for machines, is humanly impossible.

An important aspect about modern NLP techniques is the need for data: the state-of-the-art models are all composed of Deep Neural Network architectures, which generally demand a lot of data for training. In opposition to the scenario of big data presented, restricted domains or languages other than English suffer from lack of annotated data, such as datasets for specific tasks.

Effort has been made to overcome the shortage of labeled data: first approaches report to data augmentation strategies [15], relying on methods to generalize from small sets of already annotated data; other approaches treat it as a topic modeling problem, applying unsupervised methods to create clusters, further labeling them with classes of interest. However, humans typically performed the labeling step, which can be a problem because the interpretation of clusters is often challenging, and a manual labeling error would affect large amounts of data.

The historical context presented helps explain the growing interest in the field of Low-Resource NLP [13, 5], which addresses traditional NLP tasks with the assumption of shortage of data availability. Some approaches to this family of tasks propose semi-supervised methods, such as adding large quantities of unlabeled data to a small labeled dataset [26], or applying cross-lingual annotation transfer learning [2] to leverage annotated data available in languages other than the desired one. Other approaches try to eliminate the need for annotated data for training, relying, for example, on pre-trained *task-agnostic* neural language models [25], which may be used as language information sources, as well as representation learning models [17] for word, sentence or document classification tasks. However, problems related to big texts are still challenging for these models: their performance definitely worsens when their text input is as large as a congress member speech in the Parliament may be.

Therefore, it is in our interest that complex documents, like those produced by state agencies, the Parliament and the government, can be processed by machines. Applying algorithms that can make the *noisy* information more understandable, and presenting it in a visual and simple way can, thus, help in an information overload scenario.

In addition to the aforementioned issues, our project is also guided by the following facts: *(i)* the Portuguese Language is a low-resource language[1]; *(ii)* the documents produced by state agencies are complex and extensive, thus, *noisy*; *(iii)* the documents produced by state agencies contain important information for society; *(iv)* the documents produced by state agencies are publicly available.

1.2 Objective

Following these guiding facts, this project aims to design, build and evaluate a system capable of automatically extracting relevant information from given documents produced by the Portuguese Parliament, based on combined NLP techniques. It also seeks to develop methods that can deal with the low-resource scenario in which the Portuguese language fits. We present our sub-goals to be achieved in order to meet the objective:

1. Define a system architecture to select and organize the necessary methods.
2. Implement a minimally viable prototype of the system.
3. Evaluate results of the system applied to chosen databases.

Although the system may be implemented for different domains, in this work we choose the Portuguese Parliament.

1.3 Manuscript Outline

The rest of this document is organized as follows: Chapter 2 discusses concepts, background and more related work for helping the reader understand Low-Resource NLP systems; Chapter 3 introduces, defines and explains our proposed system, describing each of the fundamental modules of its architecture, as well as presenting technologies used to implement it; Chapter 4 presents results achieved by our prototyped system and evaluates them; finally, Chapter 5 discusses the results achieved and presents the authors' thoughts for future work, evaluating the possibilities of turning this system into a product.

[1]English is used by 63.1 % of the population on the internet, while Portuguese, for instance, is only used by 0.7 %. Statistics available at `https://w3techs.com/technologies/overview/content\_language`.

2 THEORETICAL FOUNDATION AND STATE-OF-THE-ART

2.1 Data Collection

As we presented earlier, the amount of public data being produced is growing, and although this brings new opportunities, it also pushes for better and more efficient techniques of Data Collection. When dealing with unstructured data, like text or sound, the challenges are even greater. Their lack of standardization and their variation in format make the Data Collection process more costly, requiring some extra steps to prepare the data. Here we present some of the more commonly used techniques for gathering publicly available data, Web Scraping and Web Crawling.

Web Scraping is a technique [41] for extracting data available in the World Wide Web (WWW) and saving it to a database, usually done with the use of Hypertext Transfer Protocol (HTTP) requests to web services. Since the WWW is where most public data is being shared, Web Scraping is becoming widely adopted for big data collection.

This technique involves both acquiring and formatting the data available in the Internet. For acquiring resources, HTTP requests are sent to the desired website. This resources can vary in format, the most common ones being web pages built in HTML, data in XML, JSON objects or multimedia data, like images, videos and audios. Once data is downloaded, the process of parsing, cleaning and structuring the data begins, until it finally outputs the data from the Web structured into a database.

The process of Web Scraping can be accomplished either manually or automatically by a bot or web-crawler.

2.2 Natural Language Processing

Natural Language Processing, commonly referred as NLP, is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between com-

puters and human language. [existe NLP feita pelo cérebro humano. aqui se trata do processamento feito em computadores. colocar footnote desambiguando?] The importance of this area is evident, as its algorithms and systems can grab aspects of human communication. These systems, in turn, aim to complete useful tasks that involve language, such as chatbots that can automatically answer client requests, algorithms that search for content in the web, like Google Search, and personal assistants, like *Alexa* and *Siri*. Research is also being carried out on complex, modern tasks, like fake news detection and hate speech recognition.

The origins of NLP remount to the Turing Test [36], in which Alan Turing discusses the question 'Can machines think?'. Since then, NLP has been a growing area of computer science, first in the form of manually coded rules. It further evolved into the field of statistical NLP, leveraging a revolution in the field of linguistics [22], within the internet and technological breakthrough. Among other authors [1] [31], we argue that some of the main challenges of computer science lie in the NLP field. Currently, the best results reported in NLP are those from Neural Networks, which we detail below.

2.2.1 Deep Neural Networks for NLP

The current approaches resort to deep learning techniques to achieve high levels of information abstraction. The use of deep networks - networks with many processing layers - frequently requires some transformations, linear or non-linear. The increasingly common use of deep learning occurs in line with the Big Data context mentioned above: there is more data available, thus, training networks with many layers is made possible

On the hardware side, the evolution of deep learning is associated with the development of Graphical Processing Units (GPUs) technologies. Many steps of the deep learning can be executed in parallel, and GPUs evidently help in this task. Also, the evolution of cloud computing is another factor that strongly contributed to the breakthrough that happened in the NLP field.

2.2.2 Transformers - Attention is all you need

Within Deep NLP, processing was mainly done with Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs). Although these neural architectures brought improvements to the previous methods, they lacked general language understanding, as they could only process text sequentially, while in human language, the terms

are often interdependent.

In 2017, the Transformers [37] were introduced, proposing a self-attention model: a model that relied solely on attention mechanisms to learn. Attention mechanisms work by assigning a relative relevance to each term of the text, based on every part of it. By definition, self-attention, also called infra-attention, is a mechanism that relates different positions of a single sentence in order to represent this sentence. These models ended up offering more possibilities of using general-purpose methods with previous language understanding.

2.2.3 BERT and RoBERTa

BERT stands for *Bidirectional Encoder Representations from Transformers*. It is an encoder created by Google in 2018 and has definitely brought one of the biggest revolutions in NLP. Its innovative aspect was mainly applying *deep bidirectional* training, as opposed to previous *shallow bidirectional* or *unidirectional* training approaches. The performance of BERT was tested in many NLP benchmarks, such as the *Stanford Question Answering Dataset* (SQuAD), the *General Language Understanding Evaluation* (GLUE) and the *Situations With Adversarial Generations* (SWAG), obtaining results equivalent to the previous state-of-the-art in all of them.

In its training step, BERT uses the *Masked Language Model* (MLM), wich turns random words in each sentence into a [MASK] token, and then tries to predict the original content of the token based on the context given by the words displayed to it; it also uses *Next Sentence Prediction* (NSP) for training, improving its general understanding of language structure.

Seven months later, RoBERTa [21] was published: a Robustly Optimized BERT training approach. By measuring the key impact of some hyperparameters, it is shown that BERT was initially under-trained, and that it could still achieve results better than the state-of-the-art models released after it. Again, RoBERTa achieves state-of-the-art results for GLUE, SQuAD and RACE [1] datasets.

Since then, BERT models have become widespread: they now power almost every single English based query done on Google Search[2], and are widely used by engineers and researchers because they need little to no fine-tuning in order to perform specific tasks.

[1] ReAding Comprehension dataset from Examinations

[2] Disponível em: https://searchengineland.com/google-bert-used-on-almost-every-english-query-342193

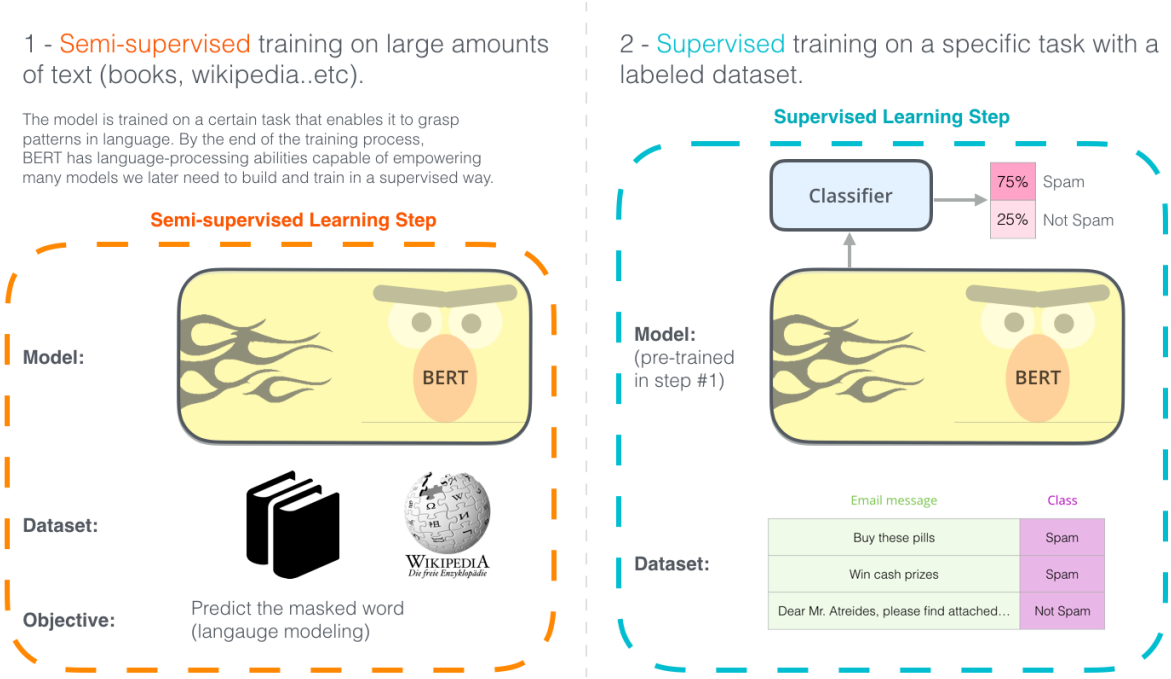

Figura 1: Google BERT on specific tasks

This is due to the Transfer Learning technique, which consists of leveraging a pre-trained model for executing a task it didn't learn in its training step (*ie.* fine-tuning the model), as shown in Figure 1.

A significant setback for Transformer-based models is related to their input size: most pre-trained Transformers cannot process inputs bigger than 512 tokens at a time. By their nature, these models pay attention to every part of the sentence, assigning numbers that indicate the relevance of the term in its context. This leads to another problem: attention cannot keep track of all information present in long texts, which makes the model perform worse. We also cite the high execution time they present, as they tend to be complex models, with millions of parameters - BERT, for example, has 334M parameters.

2.2.4 Topic Modeling

A *topic* is defined as a set of words that belongs to the same semantic field, like "school, teachers, class" or "5G, science, network". Topic Modeling can be understood as a set of techniques used to model and automatically extract distributed semantic representations from large collections of text corpora [16]. Currently, it's mostly done in a few general steps [9]: (i) embedding the documents (learned representation of words), (ii) clustering the documents in semantic similar clusters and (iii) creating topic representations from clusters.

It is a classic case of unsupervised learning in AI: given a set of unlabeled data, the algorithm finds hidden structures in it, further grouping data into automatically determined clusters. Topic modeling is especially useful when one has an enormous quantity of text documents and would like to separate them into groups that share common characteristics. For instance, it can be applied to news articles, grouping them by subject [14]. [34] measures the variation of topic distributions in 22 leading transportation journals from 1990 to 2015, and find that topics on sustainability, travel behaviour and non-motorized mobility are becoming increasingly popular over time.

2.3 Zero-Shot Learning

Much attention has been given to low data availability scenarios. Recently, an approach that has been conceptualized and developed is Zero-Shot Learning [30]. It consists of learning a classifier without having seen examples of the target class previously. This paradigm is analog to the ability humans have of recognizing an object without having seeing it in the past, just by having a high-level description of that object. For a model, it is done building an intermediate semantic layer which, instead of learning the classes themselves, helps the model to learn attributes of those classes.

It can, then, leverage its training on seen classes to predict unseen classes. This is especially useful when no training data is available, which may happen in many domains (text, voice, image) of AI applications. Some Zero-Shot approaches report results equivalent to supervised methods in the image domain [18], showing that need for training data can be surpassed.

In the text domain, [38] defines *Definition Wild* 0SHOT-TC: it aims to learn a classifier $f : X \to Y$, whereas classifier $f(.)$, however, does not have access to data X specifically labeled with class Y. We can use the knowledge that pre-trained language models already have to learn the intermediate layer of semantic attributes, which is then applied at inference time to recognize unseen classes during the training stages [40].

Standard approaches to the 0SHOT-TC task treat it as a *textual entailment* problem: given two documents d_1, d_2, we say "d_1 *entails* d_2"$(d_1 \Rightarrow d_2)$ if a human reading d_1 (named *premise*) would be justified in inferring the proposition expressed by d_2 (named *hypothesis*) from the proposition expressed by d_1 [19].

In the case of 0SHOT-TC, d_2 is the hypothesis $\mathcal{H}(l_j)$, which is simply a sentence that expresses an association to l_j. For example, for categorizing speeches given in the

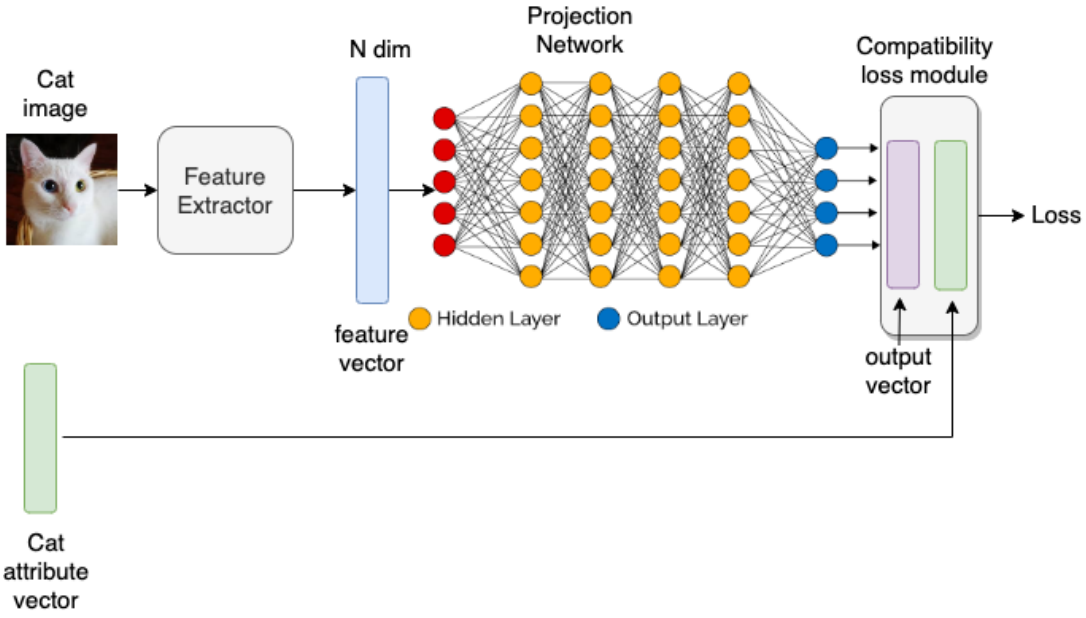

Figura 2: Zero-Shot Learning

Parliament, a label could be "**education**"and a hypothesis for it could be "This speech is about **education**". Creating the hypothesis is essential to make it understandable by a Language Model, and allows us to discover the probability $P(l_j|d_i) = P(d_i \Rightarrow \mathcal{H}(l_j))$, as $P(d_i \Rightarrow \mathcal{H}(l_j))$ can easily be inferred by a LM, using d_i and $\mathcal{H}(l_j)$ as inputs. This inference, however, is quite demanding computationally.

3 ZEROBERTO-CITIZEN

In this chapter, we aim to present the general architecture of the proposed system, along with the methodology of its development. We include technologies, requirements and implementation details of our prototype below.

3.1 Proposed System

Our system, `ZeroBERTo-citizen`, is divided into two modules: `ZeroBERTo-sentinel` and `ZeroBERTo-reporter`. The first module, `ZeroBERTo-sentinel`, has no interaction with the user, and is responsible for Data Collection and Processing steps. The second module, `ZeroBERTo-reporter`, has 1 input given by the user: *(i)* a query consisting of filters related to author, party or time interval of documents it wishes. Its general flow is shown in Figure **??**.

The first processing steps are done independently from user input. `ZeroBERTo-sentinel` automatically collects public data available by web scraping, then structures this data into an initial database. This database is given as input to the Topic Modeling module, which generates topics and feeds the `ZeroBERTo-reporter` database with this information.

Generated topics are also given as input to the Zero-Shot classification module, which then associates topics to predetermined classes. After that, it composes the Topic Modeling output with the Zero-Shot output, calculating Document to Class association. This information is then given to `ZeroBERTo-reporter` database.

Figura 3: ZeroBERTo-citizen Architecture - Level 0

The input given by the user is a query, which consists of a time interval and filters related to author, party or date of the speeches given. `ZeroBERTo-reporter` receives the user query, it searches in the final database for data that corresponds to filters given and feed the Graphical Interface, presenting it to the user.

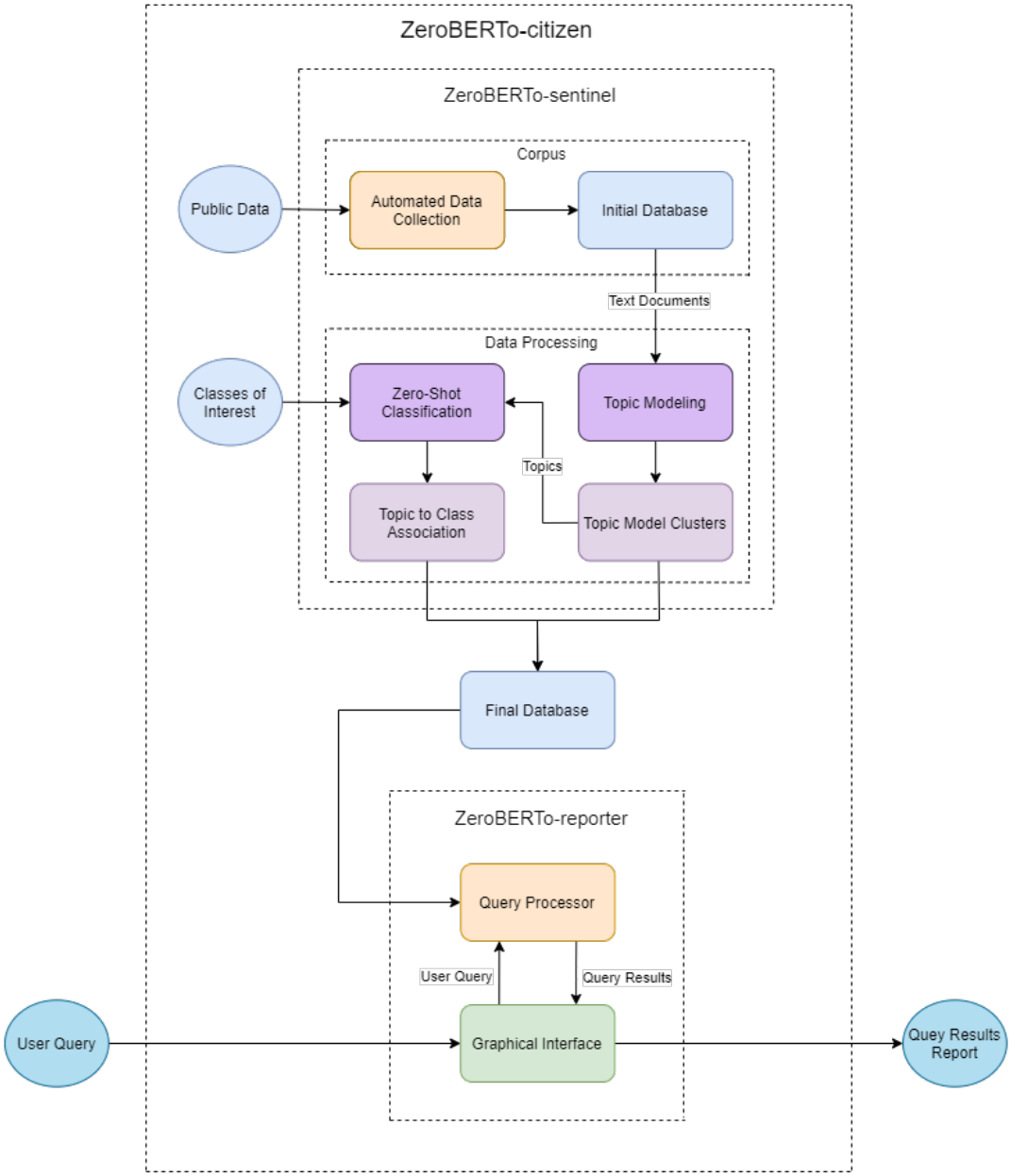

Figura 4: ZeroBERTo-citizen Architecture - Level 1

3.2 Technologies

For the development and implementation of this project, we used some technologies for Data Collection, Cloud Computing and Machine Learning, which are described below.

3.2.1 Python 3.9.2

Python is a programming language commonly used worldwide. It is a versatile tool both for object-oriented programming and for structured programming. In our context, Python has shown to be a versatile and effective choice to structure our data and to process information available in the internet, and we will be using its version 3.9.2. Since it´s highly utilized for data science, Python also has modules implementing a lot of Machine Learning models. Also, it has an enormous online community, with extensive guides and willing to help experts. Finally, many common Python libraries offer an intuitive way for data mining and visualizing.

3.2.2 Models

Below, we present the Natural Language Processing models used in this project. We justify our choices in the context of the Portuguese Parliament, explaining and detailing why these models were chosen.

`ZeroBERTo` - Combined Model

`ZeroBERTo` is a combined model specifically developed for Low-Resource NLP that works in two steps. It receives as input: *(i)* a set of unlabeled documents; *(ii)* a set of classes. It has two outputs: *(i)* a topic representation of each document d_i (middle output); *(ii)* a probability vector Θ_i of entailment between each document and every class. Technically, it also receives as input two Models: one for the first step, another one for the second step. This choice comes in handy because `ZeroBERTo` does not need labeled data, as is the case of the minutes we work with.

`ZeroBERTo` leverages an unsupervised topic modeling step, using it as a compress representation of documents data. With the Topic Model trained, instead of analyzing the relation between document d_i and class c_j, it determines the entailment between the learned topic representation $\Omega_{TM}(d_i)$ of each document and each class c_j. Topics found are given as input to the second Language Model to infer entailment probabilities. It then solves the 0SHOT-TC task by calculating a compound conditional probability for

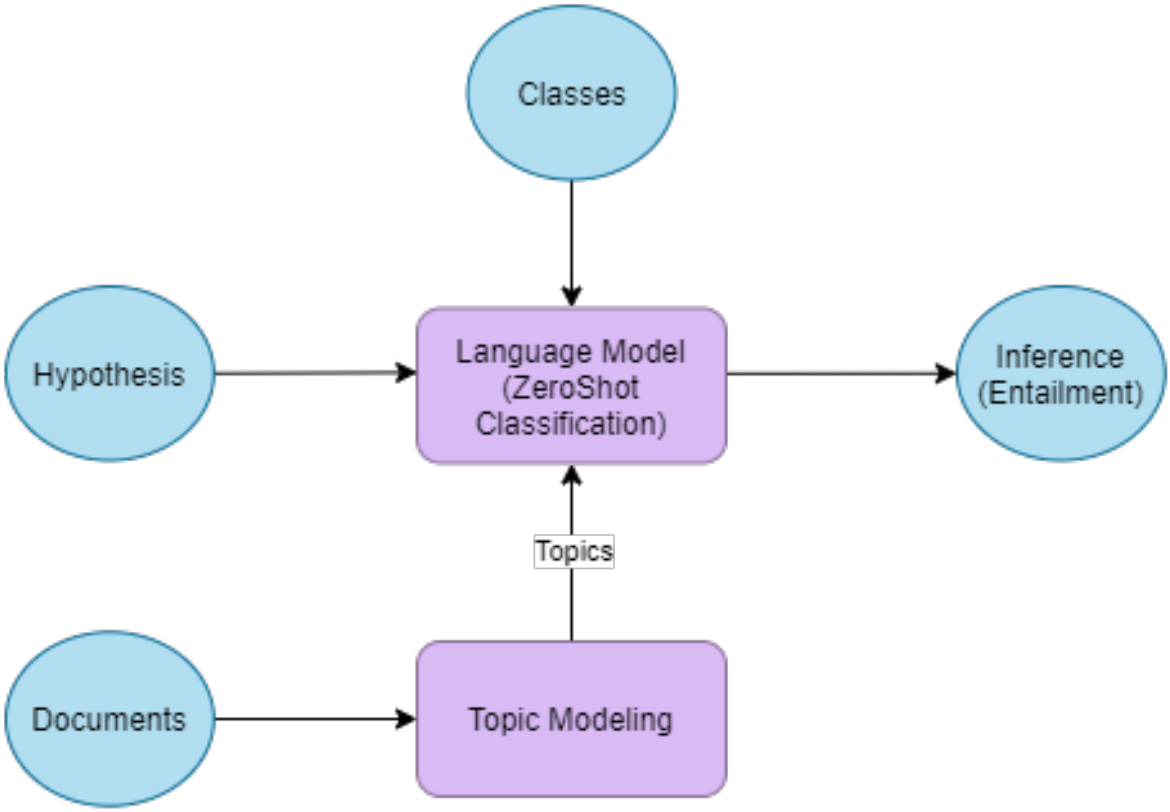

Figura 5: ZeroBERTo's two steps

each class c_j to determine the entailment vector $\Theta_i = (\theta_i^1, \theta_i^2, \ldots, \theta_i^m)$. Classification is then carried out by selecting $argmax_{(c \in \mathcal{C})}(\Theta_i)$.

`BERTopic` - Topic Model

`BERTopic` is a topic modeling algorithm [12] that receives as input: *(i)* minimum topic size (how many n-grams are needed to delimitar a new topic); *(ii)* n-grams range (which n-grams should be considered for topic representation); *(iii)* top n words (how many words/n-grams will be used for topic representation). `BERTopic` uses an Embedding Model as its first step, then reduces dimensionality of the embeddings and further groups the resulting embeddings into semantically similar clusters. It is different from traditional topic modeling techniques, like Latent Dirchlet Allocation (LDA), in the sense that it provides a continuous topic modeling, opposed to discrete ones. We show `BERTopic` general architecture along with its specific implementation for this work in Figure 6.

After the clustering step, it applies *class-TF-IDF* to extract meaningful information from the topics found in the documents, selecting the most relevant n-gram candidates. Once trained, it can be used to infer the topic representation of each document d_i, which is a vector $\Omega_{TM}(d_i)$ that contains the respective probabilities of the document d_i belonging to topic t_k.

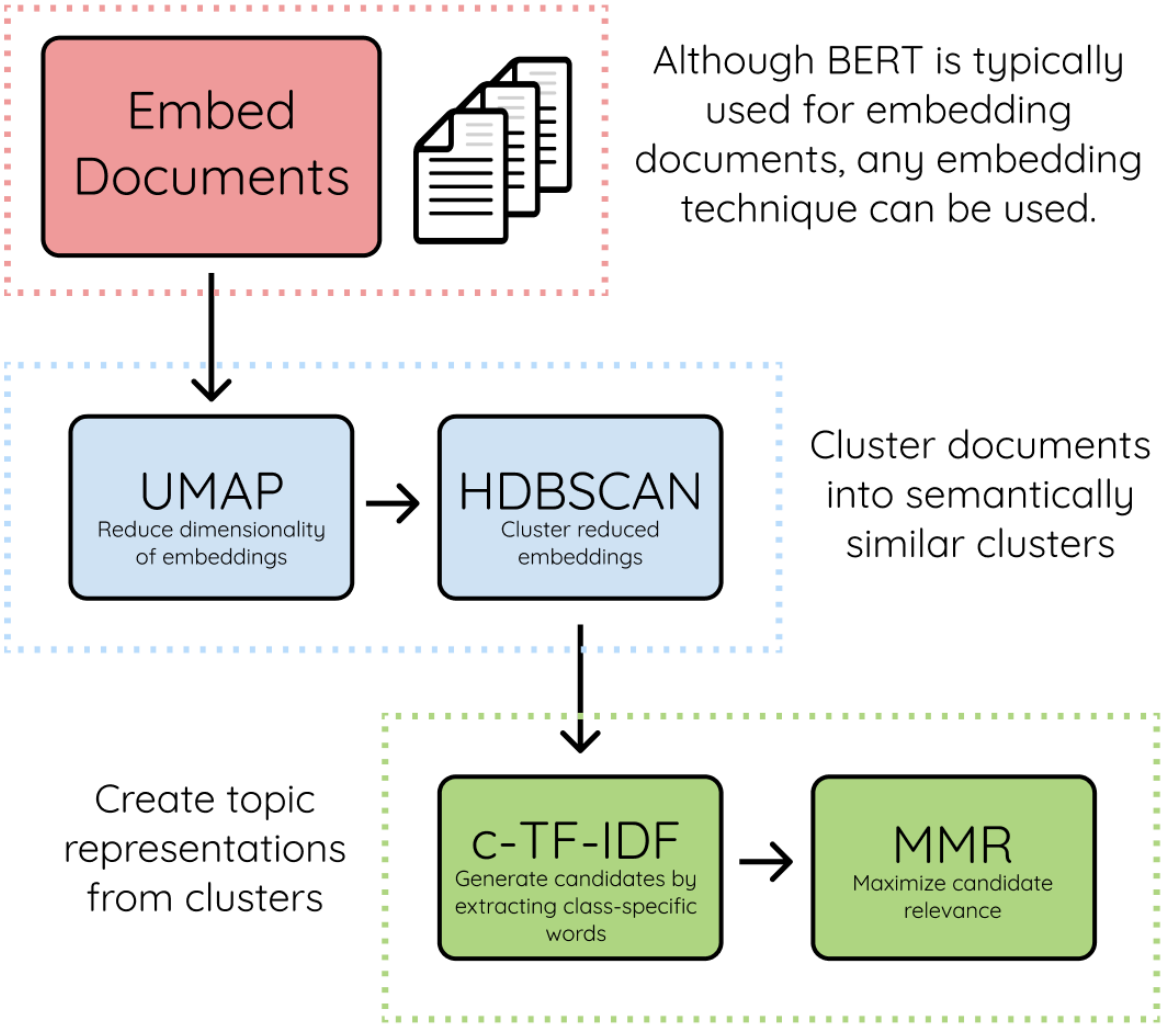

Figura 6: BERTopic implemented architecture

`M-BERT-large` **- Embedding Model 1:** The topic modeling step (implemented with `BERTopic`) needs an Embedding Model. In this work, M-BERT-large (Multilingual BERT) was used for this task. It is available on hugging face[1], and texts larger than the maximum size of M-BERT (512 tokens) are padded.

`XLM-RoBERTa-XNLI` **- Embedding Model 2:** The second processing step, Zero-Shot classification, needs another Embedding Model. We choose `XLM-R`, which is short for XLM-RoBERTa-large-XNLI, available on hugging face[2], which is state of the art in Multilingual 0SHOT-TC. It is built from XLM-RoBERTa [6] pre-trained in 100 different languages (Portuguese among them), and then fine-tuned in the XNLI and MNLI datasets (which do not include the Portuguese language). It is already in the zero-shot learning configuration described in paper [38] with template hypothesis as input. Texts larger than

[1]Available at: https://huggingface.co/bert-base-multilingual-cased
[2]Available at: https://huggingface.co/joeddav/xlm-roberta-large-xnli

the maximum size of XLM-R (512 tokens) are padded.

As reported by [20] and because it is not fine-tuned on Portuguese, this model has natural limitations of multilinguality. However, in this work we seek to address a Low-Resource scenario, so we overlook these limitations.

3.2.3 Libraries

Machine Learning is used in many different tasks with Python. As mentioned before, Python has, thus, a lot of packages and libraries developed to address these tasks. Next, we present libraries used throughout this project.

Selenium: Selenium is an umbrella project for a range of tools and libraries that enables and support the automation of web browsers. It provides extensions to emulate user interaction with browsers, a distribution server for scaling browser allocation, and the infrastructure for implementations of the W3C WebDriver specification that lets you write interchangeable code for all major web browsers. There's a Python API for this library, which we use. It will serve for automatic Data Collection since it provides ways to interact with JavaScript elements on websites.

BeautifulSoup: For easily extracting and manipulating the information present in the resources extracted with Selenium, we used BeautifulSoup, a library to extract what is relevant from the pages' HTML code. With this library it is possible to parse specific elements of the HTML code as the programmer wishes. Together with Selenium, Beautiful Soup [**Beautiful Soup**] will be used for pulling data out of HTML and XML files. It provides idiomatic ways of navigating, searching, and modifying the parse tree of these files.

Pandas: Pandas [**Pandas**] is a fast and powerful open source data analysis and manipulation tool built on top of the Python programming language. It will serve as a means of preparing the data for the models and generating the resulting databases.

RegEx: Regular Expressions, or RegEx, are essentially a programming language for finding patters in strings (text). They are included in Python and available through the re module, that allows better text manipulation, being very useful for our data cleaning needs.

Transformers: Most NLP applications nowadays use Attention Models (Transformer-based) detailed in section 2.2. A common way for implementing them is with the library Transformers, which has most state-of-the-art pre-trained language models for easy im-

porting and use.

3.2.4 Google Colab (Cloud Computing)

Google Colab is a free tool which enables developers to work remotely and simultaneously on a single piece of code. It also provides computing power directly from Google servers. Colab notebooks allow us to combine executable code and rich text in a single document, along with images, HTML, LaTeX and more.

3.3 System Requirements

To define the system requirements, the first step is showing the stakeholders involved in the project and then the system's Use Cases that were mapped as Functional Requirements. At last, the non-functional requirements will be explained.

3.3.1 Stakeholders

Since the final product of this project is still open to possibilities, this section contains a partial list with the main stakeholders.

Portuguese journalists and communicators: The project main stakeholders are the Portuguese journalists and communicators that are looking for information on the activities of the Portuguese parliament and it's quantitative analyses provided by the system.

Academics interested in Portuguese political activities: Some academics that might be interested in studying and analysing Portuguese political activities can benefit from the system output and database.

Portuguese people in general: Since promoting public transparency is one of the main goals of this system, one important project stakeholder is the Portuguese people.

3.3.2 Functional Requirements

In this section, the Functional Requirements are presented in the form of use cases. As mentioned before, these requirements might still change in the future, once the technical and economical characteristics of the system as a product are clearer.

Graphical Interface: To facilitate the access and visualization of the collected data and analysis provided by the system, a website with an online dashboard will be made available. This dashboard has to be user-friendly, so that is accessible to all people, and display its data in a interactive manner.

Structured Data: The database that the system will generate alone can be useful to some stakeholders. The process of data mining and data structuring generates value to them because the current available information is preprocessed and made available in a more structured and queryable manner.

3.3.3 Nonfunctional Requirements

For the Nonfunctional Requirements, this section presents a list of the main ones.

Algorithmic Transparency: Algorithmic Transparency is the principle that the factors influencing the decision made by algorithms should be made clear and visible, or transparent, to the people who use, regulate and are affected by the algorithms employed. As an open-source project that also proposes to bring more transparency to public data, it is even more relevant for our methods to be clear and open for questioning.

Usability: Since our stakeholders are generally non-tech savvy people, the systems outputs should be displayed in a intuitive way, facilitating the understanding of the analyses and interaction with the data.

Scalability: To be prepared for an increase in either the number of accesses to the dashboard, number of requests to the bot or number of data consumed and processed, the system needs scalability.

Availability: In the context of an online tool, constant and reliable availability of the system is crucial. It is imperative that the Bot be available to respond to user requests.

Data Integrity: Data integrity deals with the integrity, consistency, and correctness of the data in the application. Dealing with public data, there is the responsibility of avoiding any distortion in the processed information.

Maintainability: Our data sources get updated on a daily basis. Our systems should expect constant updating of the data, while also providing output.

3.4 Implementation

3.4.1 `ZeroBERTo-sentinel`

Data Collection Step: This step can be divided into two phases. First, we use Web Scraping techniques to retrieve all the Minutes from a Portuguese state website containing a catalog of parliamentary debates [3]. Then we enter the Data Cleaning phase, where we read the minutes text files, find relevant data and structure it in a database.

For the first phase, we use a Selenium web-crawler that enables us to interact with the website and simulate clicks to download the Minutes as .txt files and save them to a file system. Then we read the files one by one, extracting relevant data, such as: the name of the Member of Parliament (MP) who gave the speech, to which party the MP belongs, the text of the speech itself, the minutes date. For this, we mainly used Python's module for RegEx, (*re*) for searching relevant data in the file, together with Pandas, for creating and structuring the database.

Topic Modeling Step: We implement `BERTopic` as described in paper [12]:

1. we use `M-BERT-large` (Multilingual BERT) [8] as Embedding Model 1;
2. we use UMAP [24] for reducing the dimensionality of the embeddings, as taking them all in consideration would be impossible computationally;
3. we use HDBSCAN [4] for clustering embeddings after the dimensionality reduction, as it is only $\mathcal{O}(n * lgn)$ in time (which is useful because the system should be able to process a large number of documents);
4. we apply class-TF-IDF for extracting meaningful information from the clusters, representing them with most relevant n-grams.

Zero-Shot Classification Step: We implement `XLM-RoBERTa-XNLI` as described in [21] with the Transformers library. We take the zero-shot classification configuration with inputs: document d1, document d2, hypothesis. We set the hypothesis to be: "The main theme of this list of words is {}" ("O tema principal desta lista de palavras é {}", in Portuguese).

[3] Available at: https://debates.parlamento.pt/catalogo/r3/dar/01/14/02

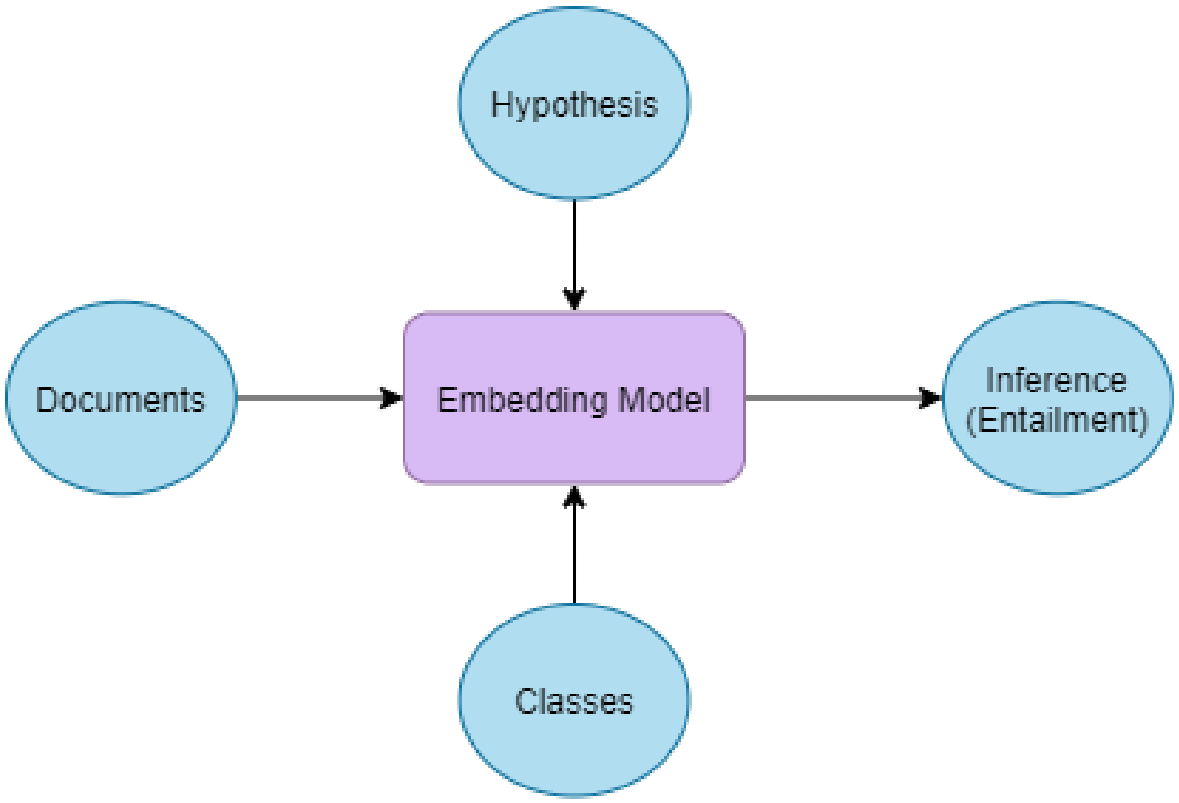

Figura 7: Zero-Shot Classification Step

3.4.2 `ZeroBERTo-reporter`

In this project, for schedule reasons, we focus only on the data processing steps, which are contained in `ZeroBERTo-sentinel`. Thus, we choose not to implement `ZeroBERTo-reporter` to its full design. Instead, we leverage data visualization libraries from Python and show results with Google Colab interface in Chapter 4.

3.4.3 System Architecture

After explaining each of the phases of data flow in the system, below we present a complete diagram of its architecture.

First, the general architecture (level 2) of the system proposed in this project is presented in Figure 8. The objective of this first diagram is to present the system in a modular way to the reader, allowing programmers to decide how to implement each individual step. Therefore, it is possible for someone to build the same system, changing parts of it in order to optimize steps of its data flow. Also, different steps of our design require parameters that will also depend on the programmer's choice, such as Topic Model hyperparameters or Zero-Shot hypothesis and classes of interest.

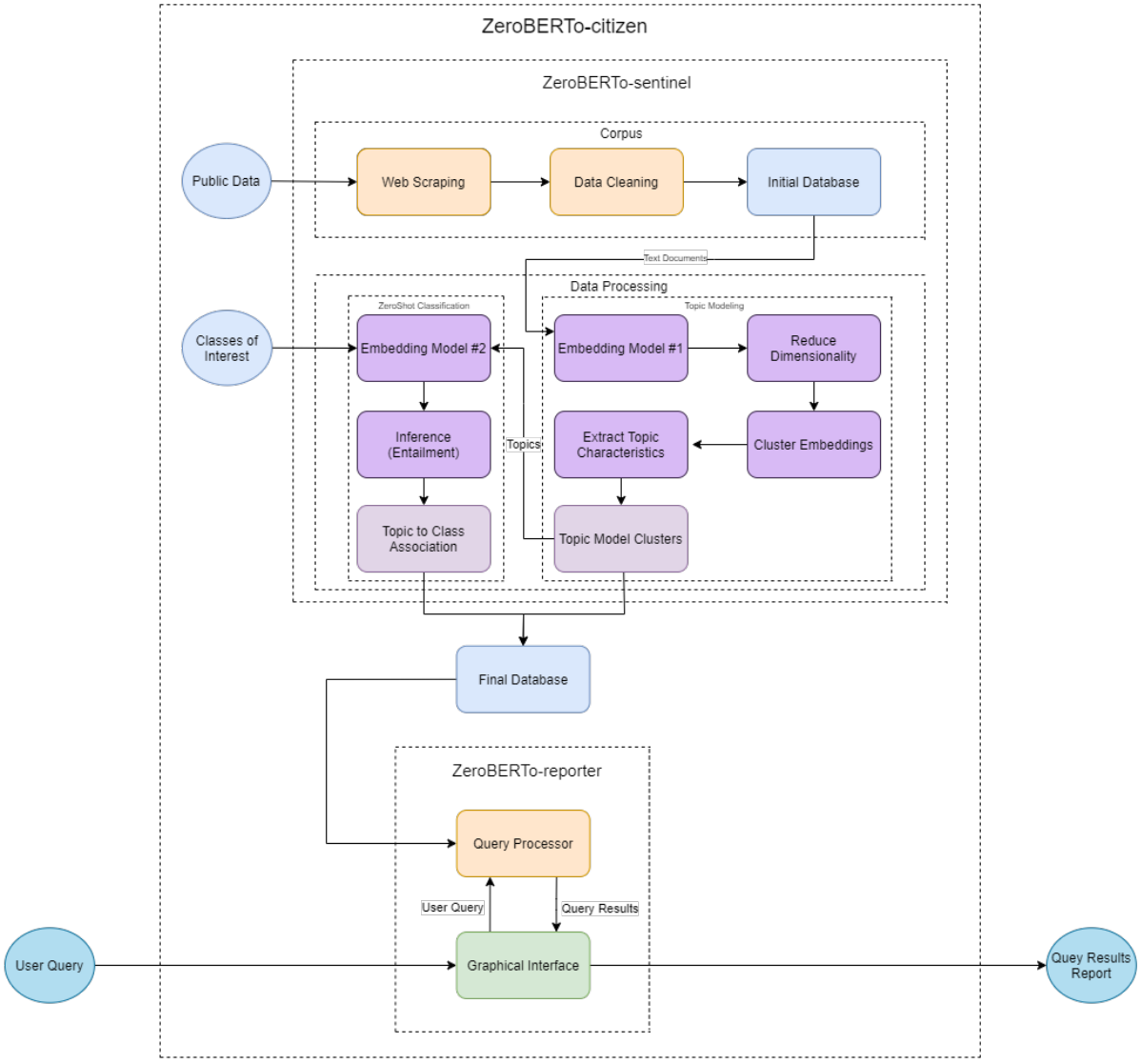

Figura 8: ZeroBERTo-citizen Architecture - Level 2

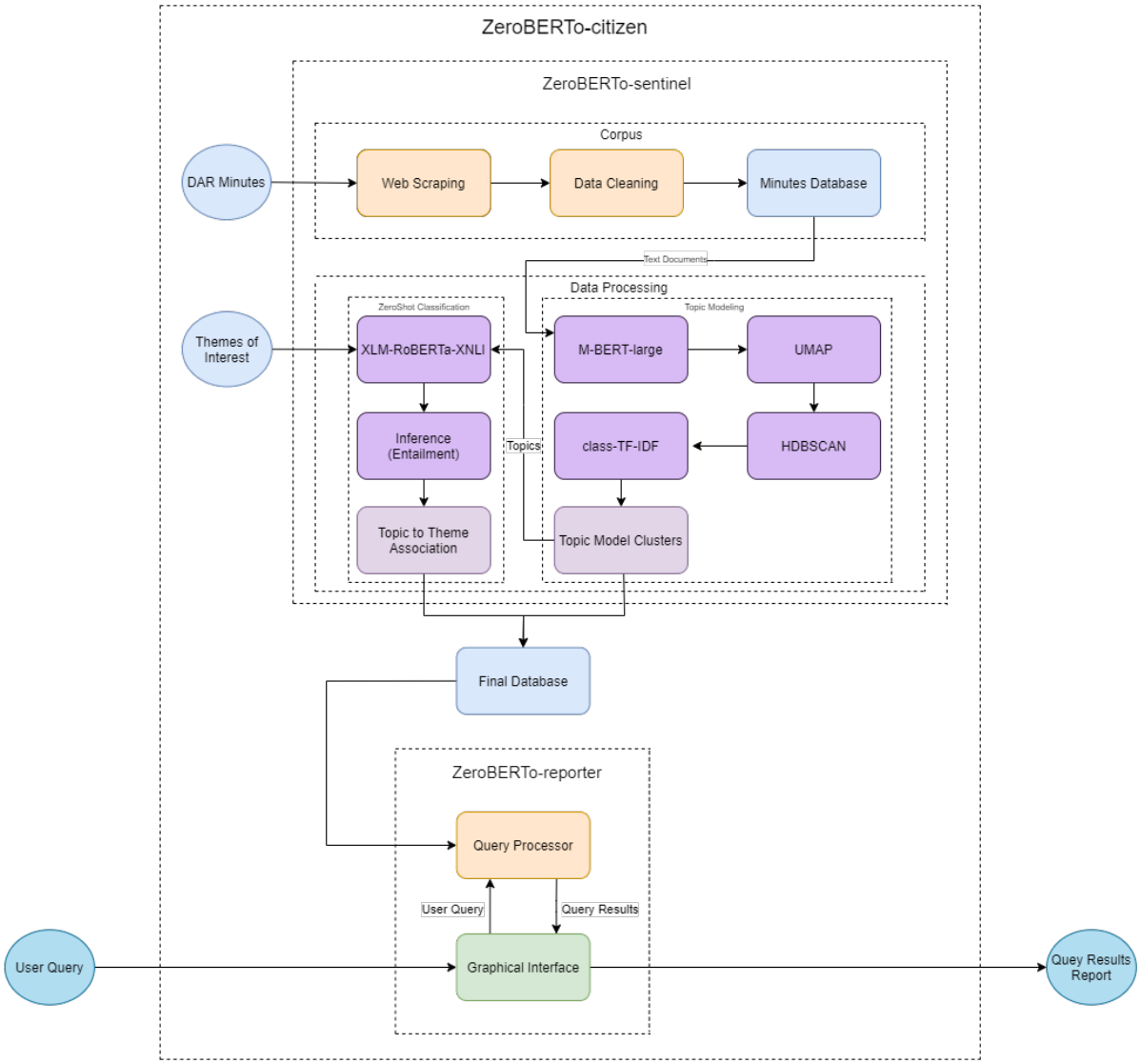

Figura 9: ZeroBERTo-citizen Architecture - Implementation

Next, we present the system architecture in Figure 8 with the specific technologies used for each phase. This gives the reader a view of where each of the models and methods were used, as well as what they are connected to.

The parameters chosen empirically for each of the phases are:

`ZeroBERTo-sentinel`: minimum topic size; n-grams range; top n words; hypothesis template; vector of classes of interest.

`ZeroBERTo-reporter`: none.

4 TESTS AND RESULTS

In this chapter, we present the tests we conducted along with obtained results. We show results from the data processing steps, explaining our choices based on intermediate stages of the data processing step.

For acquiring source data, our web scraping algorithms (detailed in 3.4.1) automatically downloads the minutes available at debates.parlamento.pt. In these tests, all data used is from the minutes contained in the time interval from 09/16/2020 to 02/25/2021. We also remove speeches with length below 500 characters, as such short speeches tend to lack semantic content related to our themes of interest, and are mostly composed of quick interruptions or questions in the Parliamentary sessions.

`ZeroBERTo-citizen` needs tuning for being applied to new domains. Specifically, its topic model hyperparameters and its zero-shot classes of interest may be changed in order to achieve higher quality results. Also, we modify the amount of data that we consider as input of the system, based on justified criteria. Below, we conduct experiments varying: *(i)* topic model hyperparameters, such as *minimum topic size* and *n-gram range*; *(ii)* the classes of interest for zero-shot classification. All tests made were using the Portuguese version of the classes listed below.

4.1 Experiment #1 - Including Legislation

We conduct the first set of experiments taking Zero-Shot classes as themes of interest by Portugal citizens. For the entailment mechanism to work, classes chosen should keep semantic similarity with our data - the speeches given in the Parliament. For that, we observe the ministerial organization of the Portuguese government. The themes chosen are:

- Legislation (*Legislação*)
- Corruption (*Corrupção*)

- Health and Quality of Life (*Saúde e Qualidade de Vida*)
- Education (*Educação*)
- Economy (*Economia*),
- Work and Employment (*Trabalho e Emprego*)
- Environment (*Meio Ambiente*)
- European Union (*União Europeia*)
- Industry and Agriculture and Commerce (*Indústria e Agricultura e Comércio*)
- Energy (*Energia*)
- Science and Technology (*Ciência e Tecnologia*)
- Tourism (*Turismo*)
- Culture (*Cultura*)
- National Defense and Public Security (*Defesa Nacional e Segurança Pública*)
- Housing and Urban Planning (*Habitação e Urbanismo*)
- Infrastructure (*Infraestrutura*)
- Justice (*Justiça*)
- Human Rights (*Direitos Humanos*)

Below we present results achieved with this set of themes. Then, we analyze and discuss some aspects of output information and try to explain the outcome of some tests.

minimum topic size = 10

Test 1: n-gram range = [1,3]

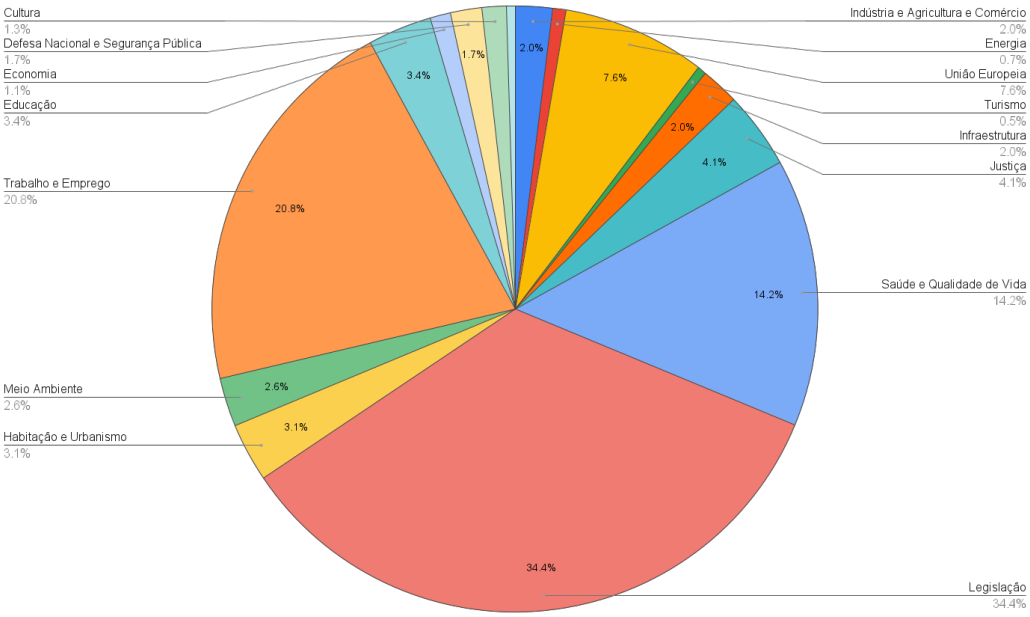

Figura 10: Test 1

Test 1: n-gram range = [2,3]

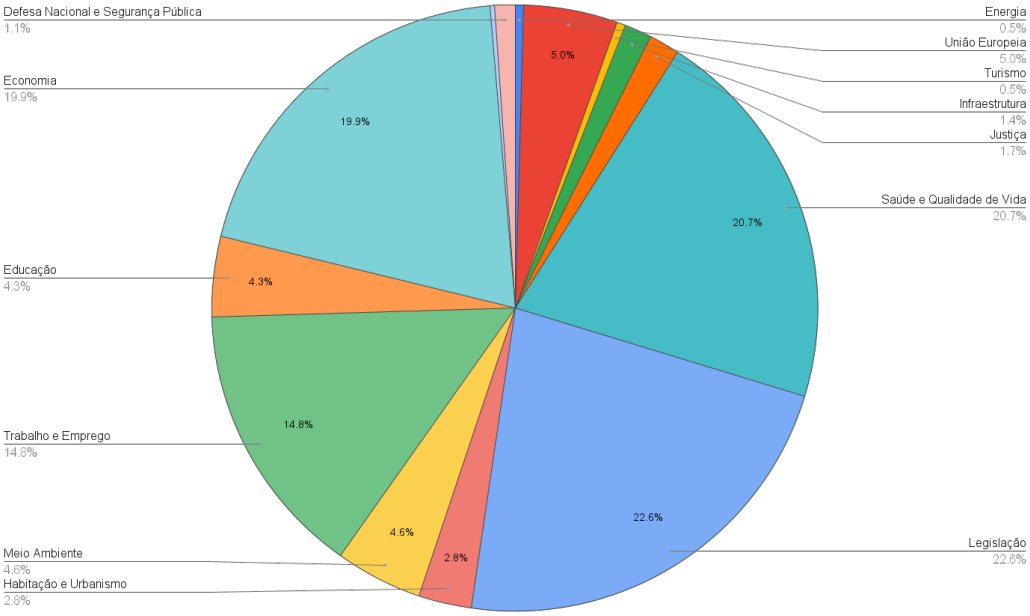

Figura 11: Test 3

minimum topic size = 15

Test 5: n-gram range = [1,3]

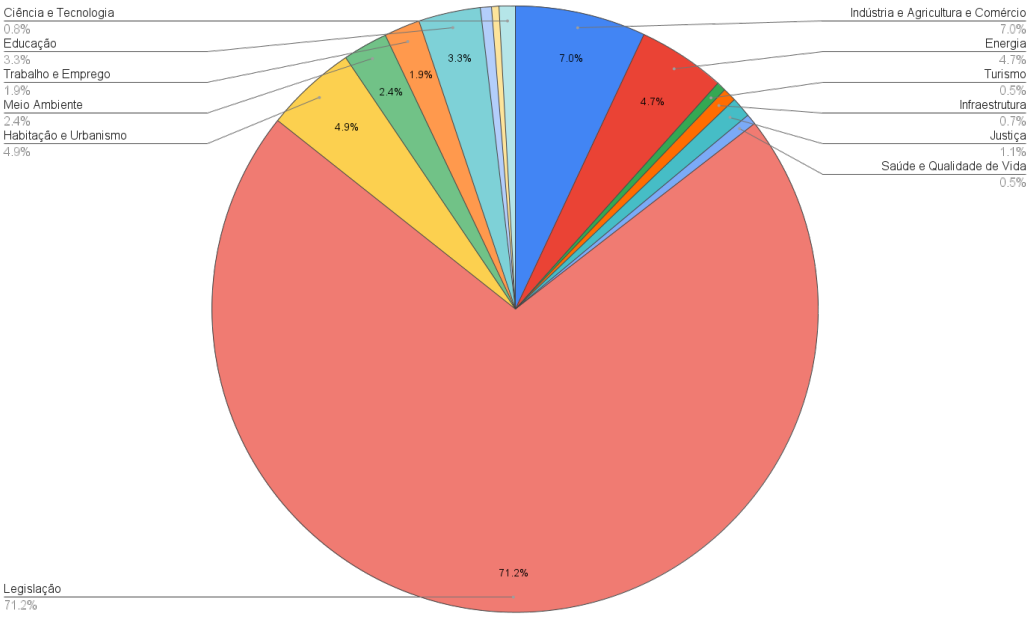

Figura 12: Test 5

Test 7: n-gram range = [2,3]

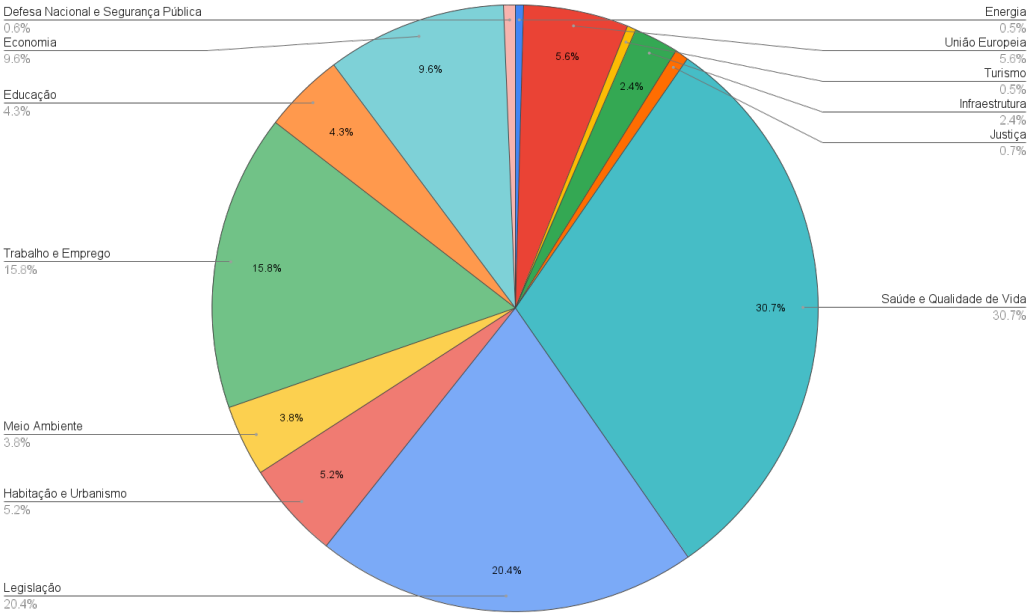

Figura 13: Test 7

minimum topic size = 20

Test 9: n-gram range = [1,3]

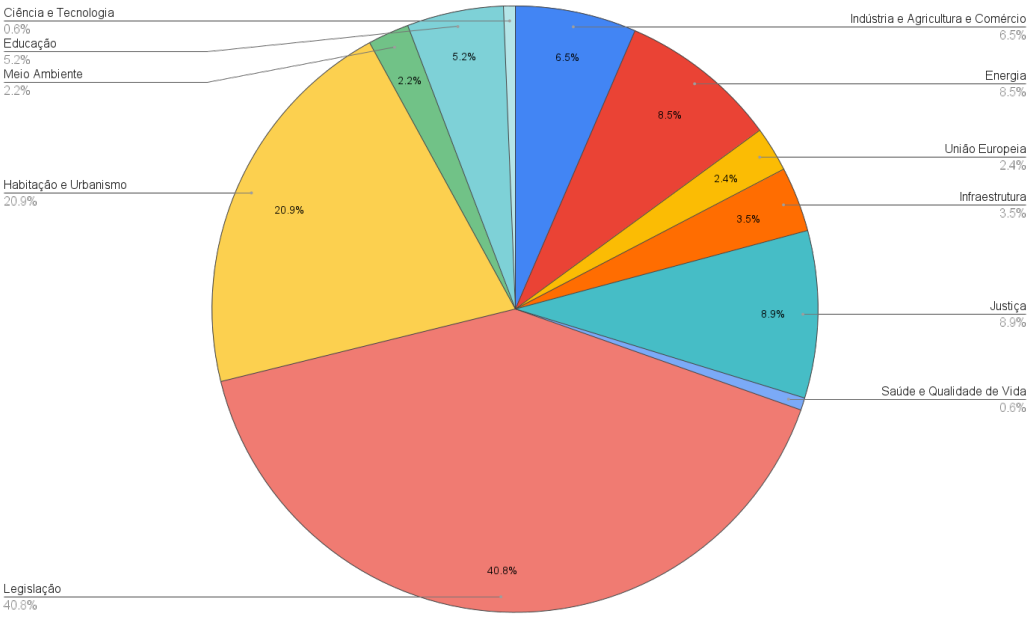

Figura 14: Test 9

Test 11: n-gram range = [2,3]

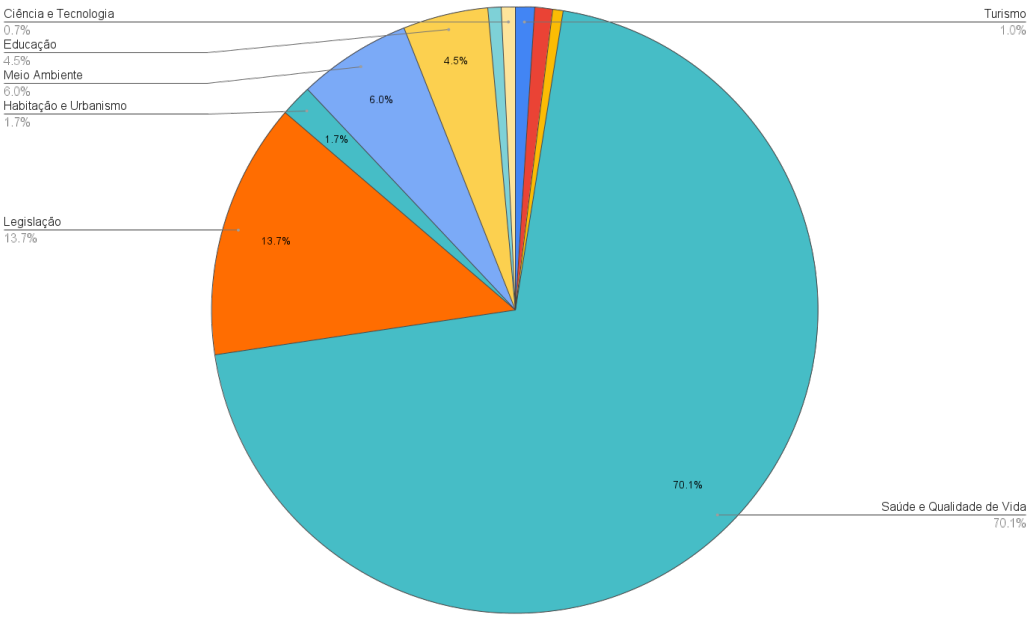

Figura 15: Test 11

Table 1 shows results from the first experiment conducted. We note that the theme

	Test 1	%	Test 3	%	Test 5	%	Test 7	%	Test 9	%	Test 11	%
Indústria e Agricultura e Comércio	77	2.0%	0	0.0%	269	7.0%	0	0.0%	250	6.5%	0	0.0%
Energia	27	0.7%	18	0.5%	180	4.7%	18	0.5%	326	8.5%	0	0.0%
União Europeia	294	7.6%	193	5.0%	0	0.0%	215	5.6%	94	2.4%	0	0.0%
Turismo	19	0.5%	18	0.5%	20	0.5%	19	0.5%	0	0.0%	40	1.0%
Infraestrutura	78	2.0%	55	1.4%	27	0.7%	91	2.4%	133	3.5%	36	0.9%
Justiça	159	4.1%	64	1.7%	42	1.1%	28	0.7%	343	8.9%	22	0.6%
Saúde e Qualidade de Vida	549	14.2%	796	20.7%	21	0.5%	1182	30.7%	25	0.6%	2701	70.1%
Legislação	**1324**	**34.4%**	**871**	**22.6%**	**2742**	**71.2%**	**785**	**20.4%**	**1571**	**40.8%**	**526**	**13.7%**
Habitação e Urbanismo	121	3.1%	109	2.8%	188	4.9%	200	5.2%	804	20.9%	65	1.7%
Meio Ambiente	100	2.6%	179	4.6%	93	2.4%	146	3.8%	85	2.2%	233	6.0%
Trabalho e Emprego	800	20.8%	569	14.8%	74	1.9%	609	15.8%	0	0.0%	0	0.0%
Educação	131	3.4%	164	4.3%	127	3.3%	166	4.3%	199	5.2%	175	4.5%
Economia	42	1.1%	766	19.9%	23	0.6%	371	9.6%	0	0.0%	0	0.0%
Direitos Humanos	0	0.0%	10	0.3%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Defesa Nacional e Segurança Pública	64	1.7%	41	1.1%	15	0.4%	23	0.6%	0	0.0%	27	0.7%
Cultura	51	1.3%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Corrupção	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Ciência e Tecnologia	17	0.4%	0	0.0%	32	0.8%	0	0.0%	23	0.6%	28	0.7%

Tabela 1: Experiment 1 - Table of Results

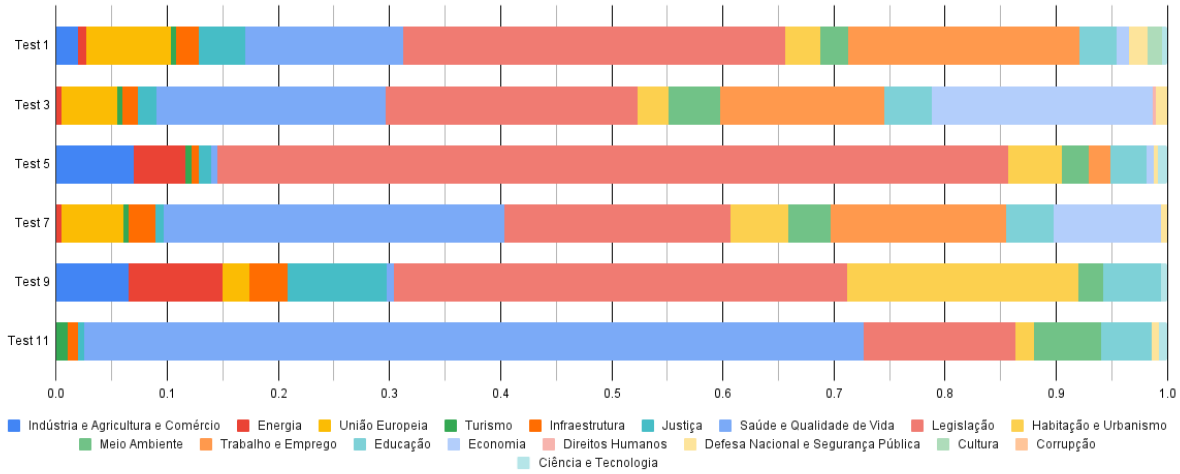

Figura 16: Experiment 1 - Theme Distribution

Legislation ends up taking a good portion of the speeches considered. When we take some speeches classified as Legislation for manually analyzing, we see that many of them contain some other theme, but formal and technical aspects of the Parliament process of Legislation are frequently discussed. This could be biasing the models.

However, in this first experiment, we see promising results as shown in Figure 16. Our manual inspection of these speeches suggest that the balance between themes shown in Test 1, Test 3 and Test 7 may be correct, as it is a characteristic of the Parliament to discuss many themes. Also, in Test 3 and Test 7, we see that **Health and Quality of Life** is highlighted, which corresponds to general expectations, as most of the minutes used were from sessions that happened during covid-19 pandemic.

It is of our interest to see how the same speeches and themes of public interest would be associated if we take the outlier for zero-shot configuration – in this case, Legislation.

	Test 2	%	Test 4	%	Test 6	%	Test 8	%	Test 10	%	Test 12	%
Indústria e Agricultura e Comércio	80	2.1%	0	0.0%	495	12.8%	0	0.0%	251	6.5%	0	0.0%
Energia	38	1.0%	41	1.1%	373	9.7%	80	2.1%	361	9.4%	0	0.0%
União Europeia	339	8.8%	221	5.7%	15	0.4%	256	6.6%	98	2.5%	0	0.0%
Turismo	19	0.5%	18	0.5%	74	1.9%	19	0.5%	0	0.0%	40	1.0%
Infraestrutura	99	2.6%	89	2.3%	191	5.0%	92	2.4%	133	3.5%	71	1.8%
Justiça	816	21.2%	279	7.2%	311	8.1%	60	1.6%	977	25.4%	306	7.9%
Saúde e Qualidade de Vida	**540**	**14.0%**	**759**	**19.7%**	**21**	**0.5%**	**1219**	**31.6%**	**25**	**0.6%**	**2810**	**72.9%**
Habitação e Urbanismo	335	8.7%	114	3.0%	**1885**	**48.9%**	588	15.3%	**1642**	**42.6%**	162	4.2%
Meio Ambiente	103	2.7%	178	4.6%	110	2.9%	151	3.9%	84	2.2%	233	6.0%
Trabalho e Emprego	1158	30.1%	1089	28.3%	120	3.1%	728	18.9%	0	0.0%	0	0.0%
Educação	131	3.4%	159	4.1%	132	3.4%	166	4.3%	199	5.2%	175	4.5%
Economia	61	1.6%	820	21.3%	38	1.0%	465	12.1%	28	0.7%	0	0.0%
Direitos Humanos	0	0.0%	10	0.3%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Defesa Nacional e Segurança Pública	66	1.7%	76	2.0%	56	1.5%	29	0.8%	32	0.8%	28	0.7%
Cultura	51	1.3%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Corrupção	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Ciência e Tecnologia	17	0.4%	0	0.0%	32	0.8%	0	0.0%	23	0.6%	28	0.7%

Tabela 2: Experiment 2 - Table of Results

So, we conduct another experiment and present results below.

4.2 Experiment 2 - Excluding Legislation

Table 2 shows results from Experiment 2. Figure 17 shows the theme distribution throughout the tests. With the outlier theme taken out, we see varying distributions of themes depending on topic model hyperparameters. Some facts should be noted:

1. **n-gram range** configuration set to **[2, 3]** tends to benefit **Health and Quality of Life** as main theme. This is seen comparing Tests 2, 6 and 10 to Tests 4, 8 and 12: in the latter, the theme has a minimum of 20% from total speeches.

2. **House and Urban Planning** is severely over-chosen in Tests 6 and 10. The manual inspection conducted on speeches classified into this theme suggest that this is due to the frequency Members of Parliament refer to the Parliament itself as "this house", for instance. This context-specific use of words like "house"and "building"are difficult for zero-shot algorithms to grasp, as in this task it is assumed that the models can leverage previous general knowledge for inference.

3. **Corruption** and **Human Rights** have no occurrences in speech classification in every scenario considered. However, it does not mean that found topics were not associated to these themes in the zero-shot step: as shown in Figure 18, there are topics associated to **Human Rights** even though the speeches themselves are not classified into that theme.

In every Test except 12, we observe balance between the same themes (approximately). This is a great result: without **Legislation**, the topic modeling step brings more

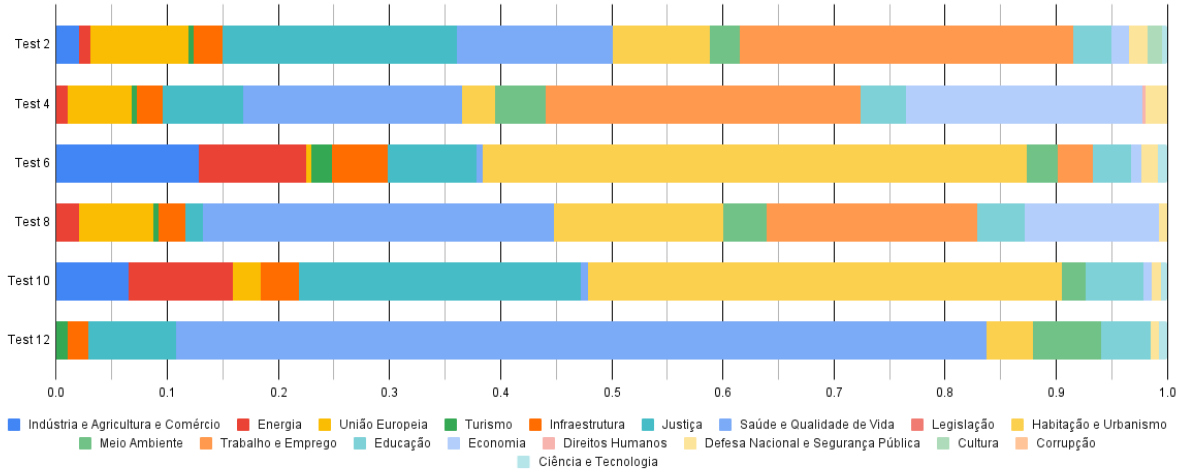

Figura 17: Experiment 2 - Theme Distribution

consistency to the output of the system. Changing topic model hyperparameters presented chaotic effects in the first Experiment. But in this one, modifying the n-grams range and the minimum topic size simply affects the prioritization of information: by controlling these hyperparameters, we can narrow down the topic analysis, potentially leaving useful information out (which seems to be the case of Tests 11 and 12), or we can widen our analysis, generating more topics but potentially confusing the zero-shot Language Model.

Finally, we present a selection of topics and themes which illustrates part of the entailment mechanism used by `ZeroBERTo-citizen`. In Figure 18, we see the 12 main topics automatically found in the documents, along with their 3 most representative n-grams. The bars represent the relative importance that the specific n-gram has inside a topic. Every topic is classified into a theme, shown above the bars. The configuration used for generating this output is the same as Test 8.

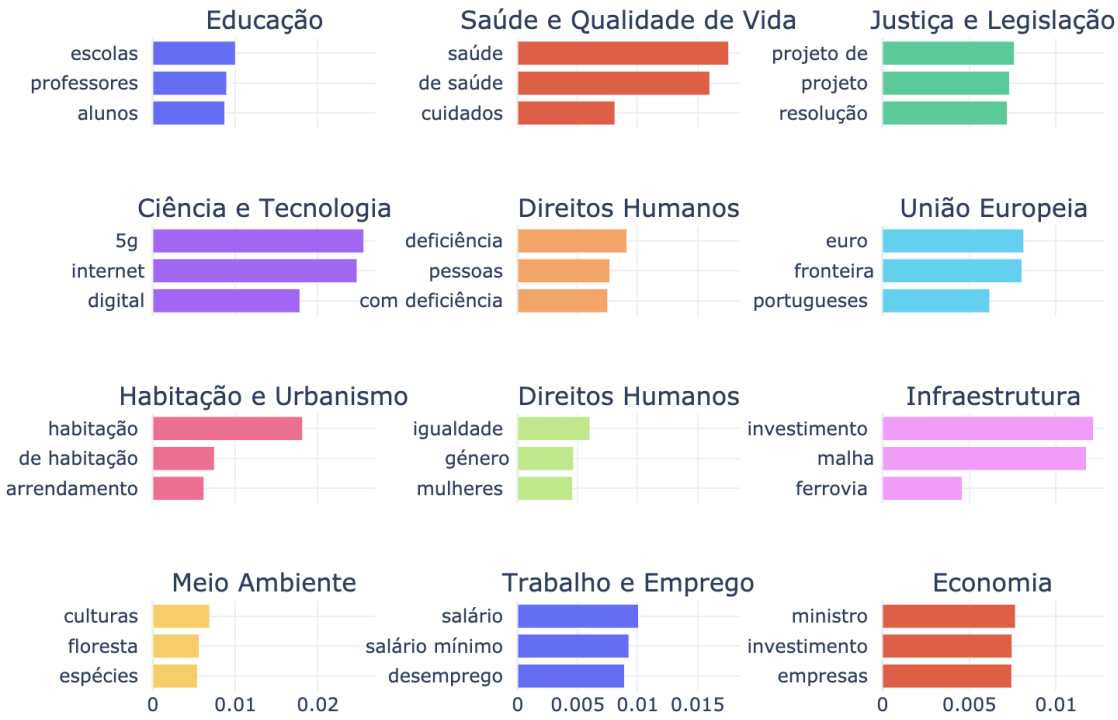

Figura 18: Main Topics and Themes

It is noteworthy that all the topics represented in Figure 18 are very coherent. Also, the association with their respective themes of interest is also very explainable. However, this topic classification does not seem to be propagated to the speech classification we aim to execute. As cited before, themes like **Human Rights**, although associated to 1 or more topics, do not get chosen in the speech classification step.

minimum topic size = 10

Test 9: n-gram range = [1,3]

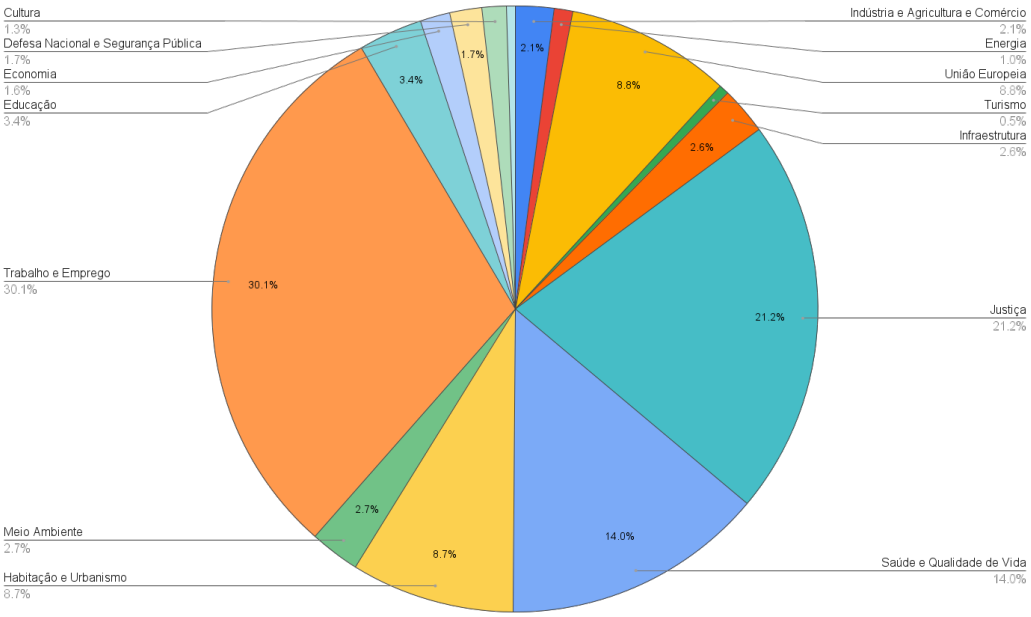

Figura 19: Test 2

Test 4: n-gram range = [2,3]

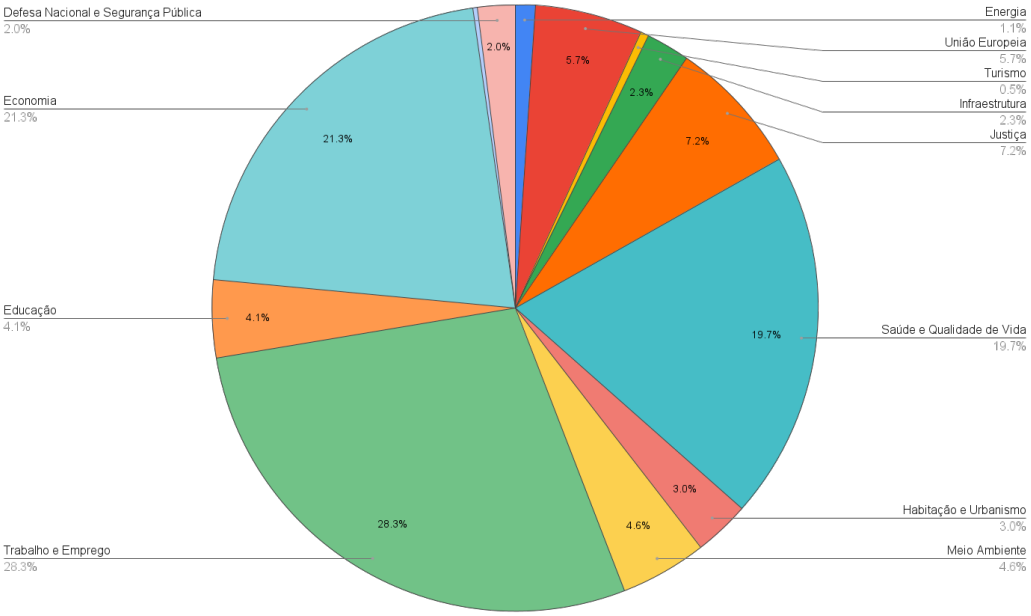

Figura 20: Test 4

minimum topic size = 15

Test 6: n-gram range = [1,3]

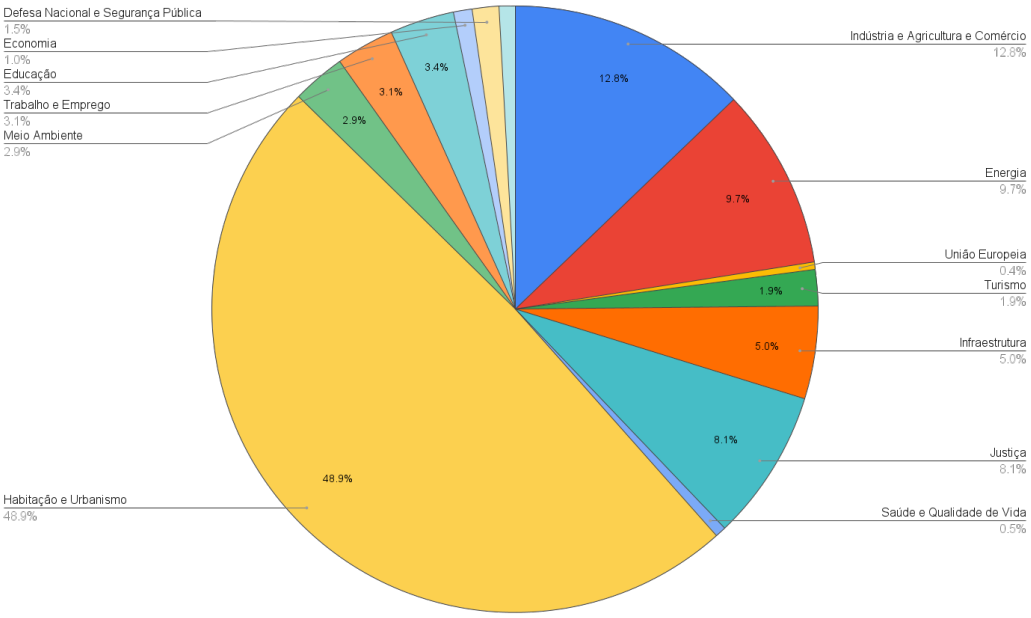

Figura 21: Test 6

Test 8: n-gram range = [2,3]

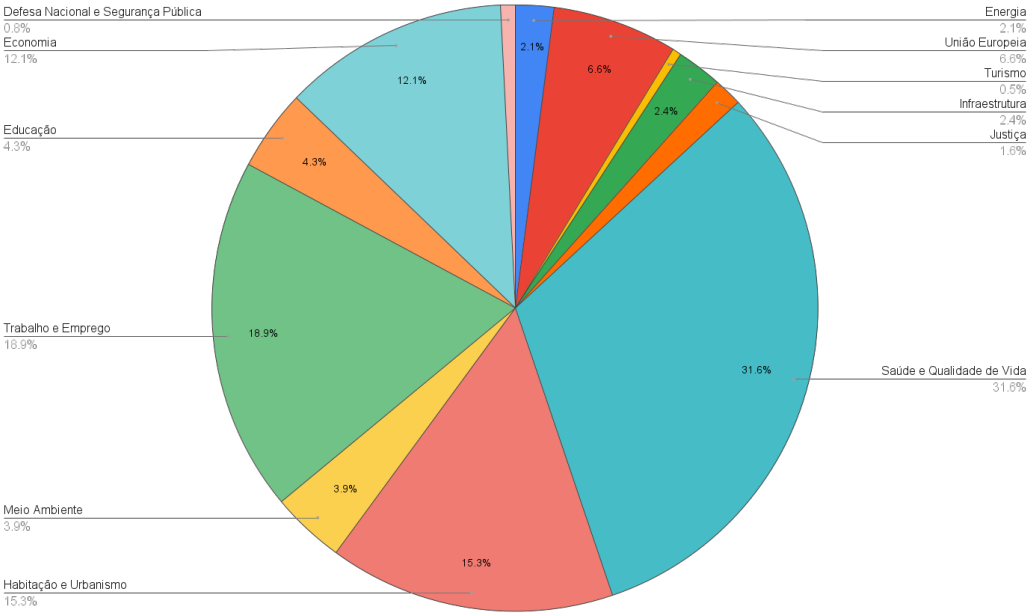

Figura 22: Test 8

minimum topic size = 20

Test 10: n-gram range = [1,3]

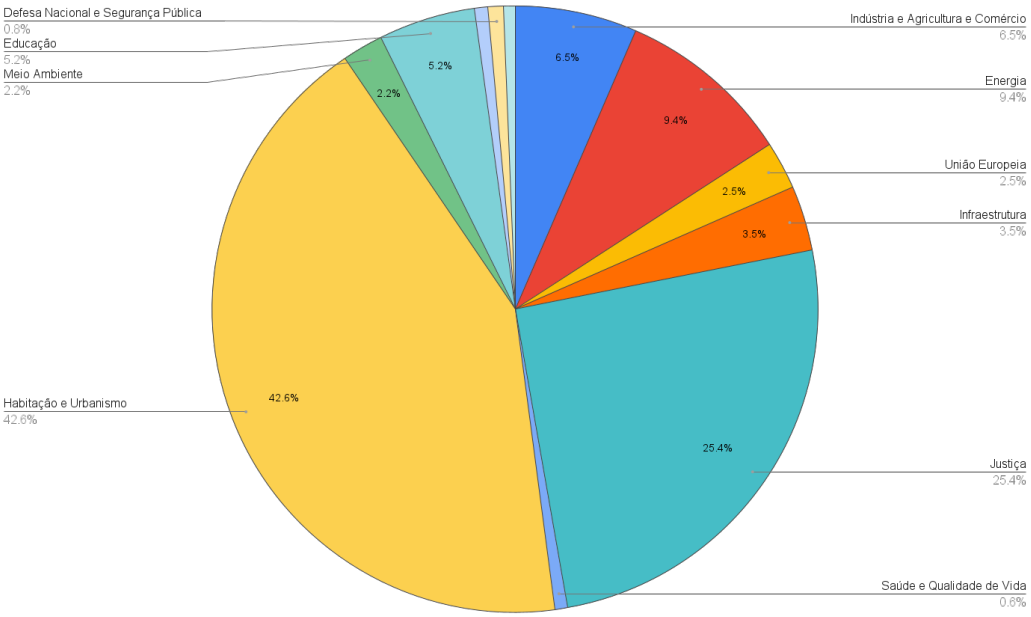

Figura 23: Test 10

Test 12: n-gram range = [2,3]

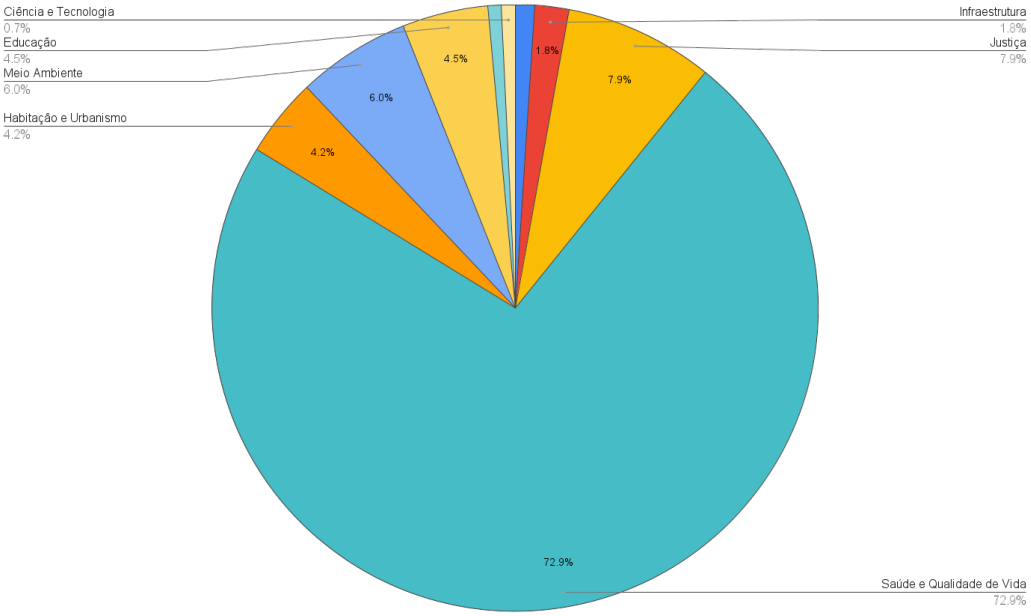

Figura 24: Test 12

5 FINAL REMARKS

We think *ZeroBERTo-citizen* addresses the issue of text classification within low-resource NLP. By applying it to the minutes of the Portuguese Parliament, we show it is possible to extract useful knowledge and highlight relevant information from sets of documents that contain long, complex texts. Everything is done with no labeled data, which makes it difficult to assess the overall performance of the system. Because of this, we chose `ZeroBERTo` model, which was previously evaluated on a benchmark dataset in Portuguese, showing it achieves F1-score up to 10% higher than zero-shot `XLM-R`, state-of-the-art for Natural Language Inference.

The results of the theme classification show `ZeroBERTo-citizen` has a lot of room for improvement in many aspects:

First, it is shown that the hyperparameter tuning of the topic model can greatly improve output information, as the right hyperparameters take non-relevant data to be deprioritized or ignored in the topic modeling step. Hyperparameter tuning brings great concern in many Machine Learning models, being even more critical in the case of complex models, which often require many hours - or days - for training. `ZeroBERTo`, however, can be taken for testing in an easier way because it is composed of two separate steps, and the hyperparameters would only affect the topic modeling training step.

Second, the hypothesis for Text Entailment in the Zero-Shot step may be changed in order to achieve more precise results. Giving the topic words as input to the Language Model to infer is much similar to a "bag-of-words" approach, which attention models like BERT are not trained for. Instead, they consider all words along with their relative importance. Because of this, we think `ZeroBERTo` uses zero-shot hypothesis in an unconventional way, and we encourage other authors to run tests for measuring the impact of changing the hypothesis. During this work, we conducted preliminary experiments on changing it, but do not present any result.

Third, the classification task was carried out by simply choosing the class with the

highest compound probability. This is a simplification that does not correspond to a real-world scenario, as each document can be associated to none, one or more classes. Research in soft metrics, for instance, can yield promising results on improving the classification.

Fourth, `ZeroBERTo-citizen` could use Explainable AI techniques to improve overall accountability of the system, which may be relevant especially when dealing with public data. We also cite causability, along with explainability, as an important factor. It is known that promoting explainability [32] in algorithmic processes has a number of advantages: giving users explanations about why certain results are achieved generates trust, while offering causal information regarding generated explanation promotes emotional security for users. In the information overload context presented, it is essential that users can trust information given by the algorithm.

LISTA DE FIGURAS

BIBLIOGRAFIA

[1] Ricardo Baeza-Yates. “Challenges in the interaction of information retrieval and natural language processing”. Em: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer. 2004, pp. 445–456.

[2] Luisa Bentivogli, Pamela Forner e Emanuele Pianta. “Evaluating Cross-Language Annotation Transfer in the MultiSemCor Corpus”. Em: *Proceedings of the 20th International Conference on Computational Linguistics*. COLING ‘04. ACL, 2004, 364–es.

[3] Kalina Bontcheva, Genevieve Gorrell e Bridgette Wessels. *Social Media and Information Overload: Survey Results*. arXiv:1306.0813. 2013. arXiv: `1306.0813 [cs.SI]`.

[4] Ricardo JGB Campello, Davoud Moulavi e Jörg Sander. “Density-based clustering based on hierarchical density estimates”. Em: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2013, pp. 160–172.

[5] Ming-Wei Chang et al. “Importance of Semantic Representation: Dataless Classification.” Em: *Aaai*. Vol. 2. 2008, pp. 830–835.

[6] Alexis Conneau et al. “Unsupervised Cross-lingual Representation Learning at Scale”. Em: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 8440–8451.

[7] Council of Europe. *Electronic democracy (“e-democracy”) – Recommendation CM/Rec(2009)1 and explanatory memorandum*. 1st. ISBN 978-92-871-6647-0. Council of Europe Publishing, set. de 2009.

[8] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. Em: *arXiv preprint arXiv:1810.04805* (2018).

[9] Adji B Dieng, Francisco JR Ruiz e David M Blei. “Topic modeling in embedding spaces”. Em: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 439–453.

[10] Daniel J Fagnant e Kara Kockelman. “Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations”. Em: *Transportation Research Part A: Policy and Practice* 77 (2015), pp. 167–181.

[11] Amir Gandomi e Murtaza Haider. “Beyond the hype: Big data concepts, methods, and analytics”. Em: *International Journal of Information Management* 35.2 (abr. de 2015), pp. 137–144.

[12] Maarten Grootendorst. *BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics.* Versão v0.9.2. 2020. DOI: 10.5281/zenodo.4381785.

[13] Michael A Hedderich et al. “A survey on recent approaches for natural language processing in low-resource scenarios”. Em: *arXiv preprint arXiv:2010.12309* (2020).

[14] Geoffrey E Hinton e Russ R Salakhutdinov. “Replicated softmax: an undirected topic model”. Em: *Advances in neural information processing systems* 22 (2009), pp. 1607–1614.

[15] Paul S Jacobs. “Joining statistics with NLP for text categorization”. Em: *Proceedings of the third conference on Applied natural language processing*. 1992, pp. 178–185.

[16] Hamed Jelodar et al. “Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey”. Em: *Multimedia Tools and Applications* 78.11 (2019), pp. 15169–15211.

[17] Yangfeng Ji e Jacob Eisenstein. “Representation learning for text-level discourse parsing”. Em: *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*. 2014, pp. 13–24.

[18] Nour Karessli et al. “Gaze embeddings for zero-shot image classification”. Em: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4525–4534.

[19] Daniel Z Korman et al. “Defining textual entailment”. Em: *Journal of the Association for Information Science and Technology* 69.6 (2018), pp. 763–772.

[20] Anne Lauscher et al. “From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers”. Em: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, nov. de 2020, pp. 4483–4499.

[21] Yinhan Liu et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. Em: (2019).

[22] Christopher Manning e Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[23] Jörg Matthes et al. ""Too much to handle": impact of mobile social networking sites on information overload, depressive symptoms, and well-being". Em: *Computers in Human Behavior* 105 (2020), p. 106217.

[24] Leland McInnes, John Healy e James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". Em: *arXiv preprint arXiv:1802.03426* (2018).

[25] Yu Meng et al. "Text Classification Using Label Names Only: A Language Model Self-Training Approach". Em: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 9006–9017.

[26] Kamal Nigam et al. "Text classification from labeled and unlabeled documents using EM". Em: *Machine learning* 39.2 (2000), pp. 103–134.

[27] James W. Prothro e Charles M. Grigg. "Fundamental Principles of Democracy: Bases of Agreement and Disagreement". Em: *The Journal of Politics* 22.2 (1960), pp. 276–294.

[28] Enrico Louis Quarantelli. "Problematical aspects of the information/communication revolution for disaster planning and research: ten non-technical issues and questions". Em: *Disaster Prevention and Management: An International Journal* (1997).

[29] Hannah Ritchie e Max Roser. "Technology Adoption". Em: *Our World in Data* (2017). https://ourworldindata.org/technology-adoption.

[30] Bernardino Romera-Paredes e Philip Torr. "An embarrassingly simple approach to zero-shot learning". Em: *International conference on machine learning*. PMLR. 2015, pp. 2152–2161.

[31] Estela Saquete et al. "Fighting post-truth using natural language processing: A review and open challenges". Em: *Expert systems with applications* 141 (2020), p. 112943.

[32] Donghee Shin. "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI". Em: *International Journal of Human-Computer Studies* 146 (2021), p. 102551.

[33] Julie Simon et al. *Digital Democracy: The tools transforming political engagement*. URL: https://www.nesta.org.uk/report/. 2017.

[34] Lijun Sun e Yafeng Yin. “Discovering themes and trends in transportation research using topic modeling”. Em: *Transportation Research Part C: Emerging Technologies* 77 (2017), pp. 49–66.

[35] Joshua Aaron Tucker et al. *Social media, political polarization, and political disinformation: A review of the scientific literature*. DOI: 10.2139/ssrn.3144139. 2018.

[36] A. M. TURING. “I.—COMPUTING MACHINERY AND INTELLIGENCE”. Em: *Mind* LIX.236 (out. de 1950), pp. 433–460. ISSN: 0026-4423. DOI: 10.1093/mind/LIX.236.433. eprint: https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf. URL: https://doi.org/10.1093/mind/LIX.236.433.

[37] Ashish Vaswani et al. “Attention is all you need”. Em: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[38] Wenpeng Yin, Jamaal Hay e Dan Roth. “Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach”. Em: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 3914–3923.

[39] Tom Young et al. “Recent trends in deep learning based natural language processing”. Em: *IEEE Computational intelligence magazine* 13.3 (2018), pp. 55–75.

[40] Jingqing Zhang, Piyawat Lertvittayakumjorn e Yike Guo. “Integrating Semantic Knowledge to Tackle Zero-shot Text Classification”. Em: *NAACL-HLT (1)*. 2019.

[41] Bo Zhao. “Web Scraping”. Em: mai. de 2017, pp. 1–3. ISBN: 978-3-319-32001-4. DOI: 10.1007/978-3-319-32001-4_483-1.