INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Topic Modeling Applied to Portuguese Tiny Text

Nicole Lopes Nunes

Master in Computer Engineering

Supervisor:
PhD Ana Maria Carvalho de Almeida, Associate Professor with Aggregation, Iscte - Instituto Universitário de Lisboa

Supervisor:
PhD Ana Rita Henrique Peixoto, Assistant Professor,
Iscte - Instituto Universitário de Lisboa

October, 2025

[This page is intentionally left blank.]

TECHNOLOGY
AND ARCHITECTURE

Department of Information Sciences and Technologies

Topic Modeling Applied to Portuguese Tiny Text

Nicole Lopes Nunes

Master in Computer Engineering

Supervisor:
PhD Ana Maria Carvalho de Almeida, Associate Professor with Aggregation, Iscte - Instituto Universitário de Lisboa

Supervisor:
PhD Ana Rita Henrique Peixoto, Assistant Professor,
Iscte - Instituto Universitário de Lisboa

October, 2025

[This page is intentionally left blank.]

This is a tiny step

[This page is intentionally left blank.]

Acknowledgment

Words cannot express my gratitude to my supervisors, Ana de Almeida and Ana Rita Peixoto, for their continuous guidance and support throughout this journey. Their encouragement pushed me to aim higher and accomplish things I once thought I was not capable of, tiny step by tiny step. I would also like to thank FCT and ISTAR for supporting this research, as well as the ISDAPPP team for their valuable insights and contributions. A special thanks to PLANAPP for providing access to the data that made this work possible.

I am deeply grateful to my colleagues for all the discussions and debates in the study room, which helped me reach conclusions that would not have been possible without them.

Finally, I would like to express my heartfelt appreciation to my family (not excluding my cat, Kit) and especially to my mother, who, despite not always understanding what I was talking about, listened whenever I needed and supported me unconditionally.

This work was partially supported by Fundação para a Ciência e a Tecnologia, I.P. (FCT) [Project 2024.07395.IACDC][ISTAR Projects: UIDB/04466/2023 and UIDP/04466/2023]

[This page is intentionally left blank.]

Resumo

Este estudo aplicou técnicas de modelação de tópicos a dados recolhidos em workshops do Centro de Planeamento e Avaliação de Políticas Públicas (PLANAPP), cujo objetivo era aproximar a ciência das políticas públicas. Esta investigação procurou extrair tópicos significativos de textos muito curtos, escritos em português pelos participantes em post-its e identificar a técnica de modelação mais adequada para este tipo de dados.

Embora não existam estudos anteriores que abordem modelação de textos extremamente pequenos, parece viável alcançar esse objetivo. Modelos baseados em embeddings têm demonstrado melhor desempenho do que abordagens clássicas em tarefas que envolvem textos curtos e conjuntos de dados pequenos.

Foram criados seis conjuntos de dados com diferentes técnicas de pré-processamento e testadas duas abordagens de modelação, Latent Dirichlet Allocation (LDA) e BERTopic. No caso do BERTopic, foram comparados dois sentence-transformers: Multilingual e Al-BERTina. As métricas clássicas mostraram-se pouco fiáveis na avaliação da qualidade dos tópicos, pois não foram desenvolvidas para textos tão curtos. Para colmatar essa limitação, foi proposto o Singularity Score (SS), desenvolvido com o objetivo de replicar o comportamento dos anotadores humanos.

As análises qualitativas e quantitativas demonstraram que o BERTopic produziu resultados mais coerentes, ainda que o LDA tenha alcançado valores superiores em métricas tradicionais, como coerência e perplexidade.

Em suma, o estudo aplicou com sucesso a modelação de tópicos em textos curtos em português, identificou o BERTopic como a técnica mais eficaz e propôs o SS como uma nova forma de avaliar a qualidade dos tópicos.

Palavras Chave: *Modelação de Tópicos*, *Texto Curto*, *Português*, *Conjunto de Dados Pequeno*

[This page is intentionally left blank.]

Abstract

This study applied topic modeling techniques to data collected from PLANAPP workshops, which aim to bridge the gap between science and public policy. The objective of this research was to extract meaningful topics from tiny, user-generated, mainly Portuguese texts written during these workshops, in post-it notes, and to find which topic modeling technique was most suitable for this type of data.

Previous studies indicate that although no work has addressed modeling over very short texts, achieving it appears feasible. Furthermore, embedding-based models have been shown to perform better than classical approaches when dealing with short texts and small datasets.

Six distinct datasets, with different preprocessing techniques, were created and tested using two modeling methods, LDA and BERTopic. For BERTopic two sentence-transformers were compared, Multilingual and AlBERTina. To evaluate the topic quality, classical metrics were employed, but did not produce reliable results. The main challenge encountered was the evaluation, as most existing metrics are not designed for tiny text. To address this issue, Singularity Score (SS) is proposed with the primary goal of mimicking the annotators behavior.

Through both qualitative and quantitative analyses, it was possible to conclude that BERTopic produced more coherent results, despite the classical method (LDA) achieving higher values in traditional evaluation metrics such as, coherence and perplexity.

In conclusion, this study successfully applied topic modeling to tiny Portuguese texts, identified BERTopic as the most suitable technique, and introduced SS as a new way to assess topic quality.

KEYWORDS: *Topic Modeling*, *Tiny Text*, *Short Text*, *Portuguese*, *Small Dataset*

[This page is intentionally left blank.]

Contents

List of Figures

List of Tables

[This page is intentionally left blank.]

List of Acronyms

BERT: Bidirectional Encoder Representations from Transformers

BoW: Bag-of-Words

CM: Coherence Measure

CRISP-DM: CRoss-Industry Standard Process for Data Mining

LDA: Latent Dirichlet Allocation

LLM: Large Language Model

NLP: Natural Language Processing

PLANAPP: Centro de Planeamento e Avaliação de Políticas Públicas

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

SLR: Systematic Literature Review

SS: Singularity Score

TF-IDF: Term Frequency-Inverse Document Frequency

[This page is intentionally left blank.]

CHAPTER 1

Introduction

1.1. Contextualization

Text mining's importance within data mining is increasing as a consequence of the online era. With almost everything available online, from books to social media posts, there is countless data from which information can be obtained. Unstructured data must be transformed into information in order to reveal underlying trends or sentiments. This helps, for example, to support decision-making by business or simply to understand the public opinion about a certain issue without having to handle it manually. Topic modeling is a task of text mining that consists of clustering documents by discovering hidden semantic patterns. Topic modeling is particularly relevant for businesses to extract themes from user-reported suggestions or issues. This enables the company to identify the main causes of these problems and determine which areas need more attention.

In this dissertation, topic modeling techniques are applied to data retrieved from Centro de Planeamento e Avaliação de Políticas Públicas (PLANAPP) [1] (Center for Planning and Evaluation of Public Policies) workshops. These workshops are part of the program *"Ciência e Política Pública: como conseguir pontes"* [2] (Science and Public Policies: How to build bridges) which aims to develop informed politics and raise scientific awareness of this need. During the workshops, participants received post-it notes and were asked to write down challenges and potential solutions to bridge the gap between science and policy. These responses were then transcribed into an Excel sheet for further analysis.

1.2. Research Questions and Goals

Defining research questions and its goals is essential to structure a study. In this Section, the research questions and each of their objectives will be defined.

RQ1: Can topics be extracted from short suggestions about public policies in Portuguese?

Goal: This question serves as the foundation for the proof of concept and represents the core objective of the dissertation. It is also a research question that serves literature review, therefore, it will be addressed through the related work analysis.

RQ2: Which topic modeling technique works best for very short-text in Portuguese?

Goal: In this question, the goal is to find the best modeling technique for the data to be analyzed. This data has specific characteristics, such as being very short-text and in a non-English language, which requires a careful approach.

[1] `https://www.planapp.gov.pt`

[2] `https://www.planapp.gov.pt/ciencia-politicas-publicas/`

1.3. Methodology

The methodology that will be followed is CRoss-Industry Standard Process for Data Mining (CRISP-DM). This process focuses on six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The first phase, Business Understanding, starts by understanding the work done in the area, the goals of the project, and its requirements. In this case, a review of the literature in Chapter 2 on topic modeling in public policies is presented. The main goal of this project is to extract topics from extremely short-text suggestions about public policies in Portuguese, Figure 1.1 presents three examples.

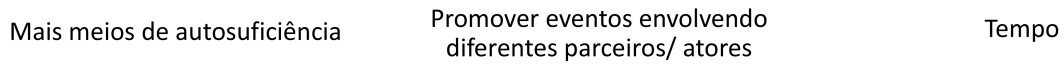


FIGURE 1.1. Suggestion Example

Secondly, in Data Understanding, which its main purpose is to acquire the data if needed and to explore it, this means describing and evaluating the quality of data, the data provided by PLANAPP and is made of short-text documents in Portuguese as it is explored in Section 3.1.

The next step is Data Preparation, where data is cleaned and prepared for modeling, these tasks encompass deleting outliers, formatting data, etc. In this project, it is important to perform the usual steps in Natural Language Processing (NLP), such as removing stop-words, correcting spelling mistakes, and turning the text to lowercase. The Data Preparation step is more detailed in Section 3.2.

Afterwards, it is the Modeling step which includes selecting modeling techniques to be applied and compared between each other. The chosen models are BERTopic with two different sentence-transformers, Multilingual and AlBERTina, and a more classical method of topic modeling, Latent Dirichlet Allocation (LDA), this process is explained in Section 3.3. Evaluation is next where it is determined if the models chosen achieve the goal set initially, qualitative and quantitative evaluation can be found in Section 3.4 and Chapter 4, respectively.

The last phase is Deployment, which can differ from project to project, varying from a report to the model being put into production. In this study, the last step is a proof of concept and the present dissertation. This step will be performed according to the plan in Figure A.1. Additionally, Figure 1.2 displays the flowchart of the methodology followed in this study.

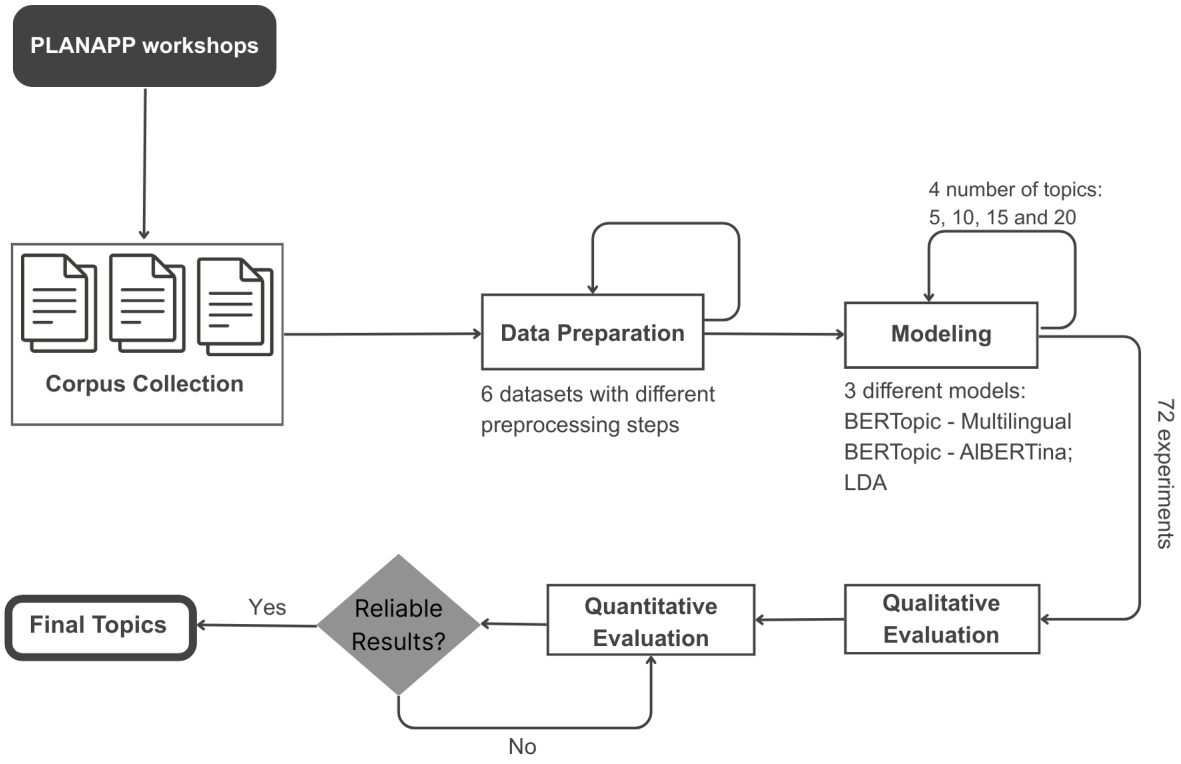


FIGURE 1.2. Methodology Flowchart

1.4. Document Structure

The document has five chapters. Chapter 1 includes a contextualization of the problem to be solved, the research questions, its goals, and the methodology. In Chapter 2 the literature related work and ways to solve similar problems are explored, as previously referred, this chapter corresponds to business understanding. Chapter 3 focuses on understanding the data, the modeling step, and making a qualitative evaluation of the results. Chapter 4 explores the quantitative evaluation using well-known metrics for topic modeling and in Section 4.7 Singularity Score is proposed as a new measurement. Finally, Chapter 5 presents the conclusions of this study, the answers to the research questions, the limitations of this work and future work.

[This page is intentionally left blank.]

CHAPTER 2

Literature Review

This chapter presents the methodology used for the literature review and its goal is to ensure a comprehensive and replicable process to identify relevant studies for our own research. The chapter first presents a general overview of topic modeling and subsequently explores a more focused review of the literature. It focuses on Portuguese and non-English, suggestions texts and the most used methods in topic modeling.

2.1. Research Methodology

The research methodology chosen was the Systematic Literature Review (SLR) based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020. This method is a way to collect, evaluate, and present systematically the findings of several studies on a specific topic [1]. The first step in an SLR is to define the topic of interest that will be the focus of the research, which is relevant to understanding how to explore and further verify if it is necessary to make the search narrower or broader [2]. In the next Section 2.2, there is an analysis of this broader search, in this case related to topic modeling. The next key step is to formulate the research questions in this review, which were:

LRQ1: What has been done recently in Topic Modeling/ Topic extraction?

LRQ2: What has been done with topics obtained from data mining about public policies or in Portuguese?

LRQ3: Can topics be extracted from short suggestions about public policies in Portuguese? - This question corresponds to RQ1.

Keywords and queries are then defined to identify relevant studies on the chosen topic in each research question. After this, the inclusion and exclusion criteria are established. Finally, an overall analysis of the documents found is conducted. In this case, each one of these last steps is explained in more detail in Sections 2.2 and 2.3, dedicated sections for the respective queries.

2.2. Topic Modeling an Overview

For a better understanding of what has been done in the last five years (2019-2024) in the main theme of NLP and topic modeling and, consequently, to respond to LRQ1, a search was performed on both Scopus and Web of Science using the query ("natural language processing" OR "NLP") AND (("topic") AND ("extraction" OR "modeling" OR "identification")). The document types analyzed were articles, reviews, book chapters, and books from Scopus, as well as articles, books, and review articles from Web of Science.

The languages included were only Portuguese and English. And finally, some author keywords such as "computational linguistics", "character recognition", and "vocabulary, controlled" (Scopus) and "linguistics" and "literature" (Web of Science) were excluded. This exclusion was made because the documents that appear under those keywords were out of the scope of the research question. This totaled 3,902 documents, 2,981 without duplicates, and 1,564 with the author keywords filtered. In Table 2.2, it is possible to observe the number of documents found in each database after applying the inclusion and exclusion criteria presented in Table 2.1.

TABLE 2.1. Inclusion and Exclusion criteria

Inclusion Criteria	Exclusion Criteria
Articles, Reviews, Books and Book Chapters Documents in English or Portuguese	Keywords: "computational linguistics", "character recognition", "vocabulary, controlled", "linguistics" and "literature"

TABLE 2.2. Documents found after exclusion criteria

Query	Data Base	Items Found
("natural language processing" OR "NLP") AND (("topic") AND ("extraction" OR "modeling" OR "identification"))	Scopus	1734
	Web of Science	2168

Figure 2.1 shows the distribution of the document types from the 2,981 records without duplicates throughout the years, and that the most published documents are articles. Furthermore, there has been a noticeable growth trend in the areas of NLP and topic modeling.

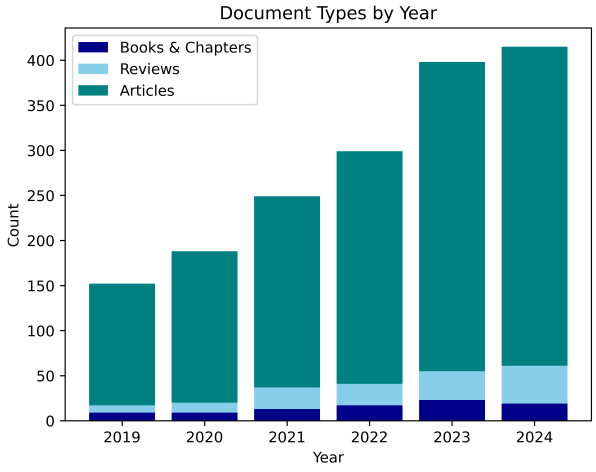


FIGURE 2.1. Document Type per Year

The graph shown in Figure 2.2 was created with the keywords of the 1,564 documents using "Vos Viewer". It represents the relationships between the keywords, where the minimum number of occurrences of a keyword was set to 8, resulting in 92 selected keywords. The minimum cluster size was set to 5, and the normalization method for performing the

clusters was LinLog/modularity. All of these parameter values were determined through exploratory analysis.

After an analysis of this graph, it is noticeable that there are a total of three clusters, and the main keyword is "natural language processing", as expected. Each of these three clusters represents a different area of NLP. The red cluster corresponds to specific methods in topic modeling, for instance: "information extraction", "text summarization", etc. The green cluster illustrates techniques used to achieve some results in NLP, such as: "bert", "word2vec", etc. The blue one represents practical applications to NLP: "social media", "covid-19", etc.

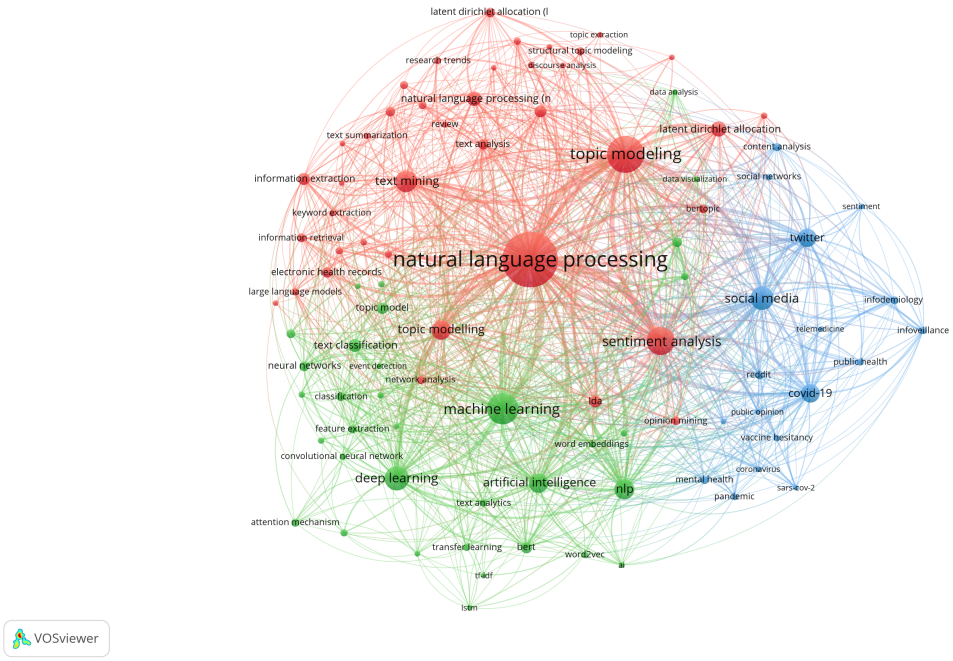


FIGURE 2.2. Keyword Relationships

2.3. Topic Modeling for Suggestions, Public Policies, and Portuguese Texts

In this Section, a specialized query was crafted to answer the second and third literature research questions: LRQ2 - "What has been done with topics obtained from data mining about public policies or in Portuguese?" and LRQ3 - "Can topics be extracted from short suggestions about public policies in Portuguese?" To do it, the search keywords have to be selected. In this case, topic modeling will be explored when it comes to public administration. It is relevant to add to the search the language in which the data is written, that is, Portuguese. Finally, a survey reference was included, as the data to be studied is obtained through it.

Using the keywords already named (Topic Modeling, Public Administration, Portuguese, and Survey), a brief search was performed in Scopus, which allowed a better understanding of the limitations caused by the use of only these keywords. The identified limitations include the fact that some references to topic extraction were not related to

its application but focused on public policies related to technology. Another flaw found was that by adding words such as survey or questionnaire, many of the documents were about surveys conducted on NLP. However, the goal was to find documents that used survey data as the primary data source. The last weak point was that some documents that dealt with the data or analyzed it were about Sentiment Analysis instead of Topic Modeling or Topic Extraction.

Bearing in mind those limitations, the initial query was polished. Therefore, a mention of approaches or techniques was added, the survey reference was removed, and finally, all documents regarding sentiment were discarded. Consequently, the final query was: (("Topic Modeling" OR "Topic Model" OR "Topic Extraction" OR "Topic Identification") AND ("Natural Language Processing" OR "NLP") AND ("Approaches" OR "Methods" OR "Techniques") AND (("Public Administration" OR "Public Sector") OR ("Portuguese Language" OR "Brazilian Portuguese" OR "European Portuguese")) AND NOT ("Sentiment")). Using the query previously mentioned and the inclusion criterion that the document type must be an Article in Web of Science, an inclusion criterion was also applied to consider only documents published up to 2024, resulting in a focus on the period between 2018 and 2024. A search was conducted on both Web of Science and Scopus, and 13 and 8 documents were found, respectively. None of the 21 documents were duplicated. Four of the documents could not be retrieved, leaving a total of seventeen. Fifteen documents remained after two of the seventeen were deemed useless to address LRQ2 and LRQ3. Three more documents were discovered through other methods, such as documents from the first query specified in Section 2.2 and snowballing, therefore, there is a total of 18 documents to be examined. Figure 2.3 illustrates this process in a PRISMA Flowchart.

2.3.1. Topic Modeling for Portuguese and Non-English Texts

In the world, there are approximately 7,000 languages, each differ significantly from the others [3]. For instance, languages can vary in word order typology, such as SVO (Subject-Verb-Object) in Portuguese and English. However, even within this similarity, differences exist, such as the placement of adjectives after the noun in Portuguese and before it in English. Additionally, lexical divergences arise, such as the meaning of a word can vary depending on the context in which it appears. Portuguese also marks the gender of adjectives, a feature absent in English. Other distinctions include the omission of pronouns and the existence of words in one language without direct translation in another [3]. These are just a few of the many differences that must be considered when dealing with natural language and its processing. With this in mind, the search included a reference to the Portuguese language. Nevertheless, among the 18 documents retrieved following the PRISMA 2020 guidelines described in Section 2.3, only one explored the textual data originally written in Portuguese. In nine of the documents analyzed, the data was written in English. Three explored the text in Chinese, while the other three were multilingual, and then three where the original language was not specified, although it is likely that

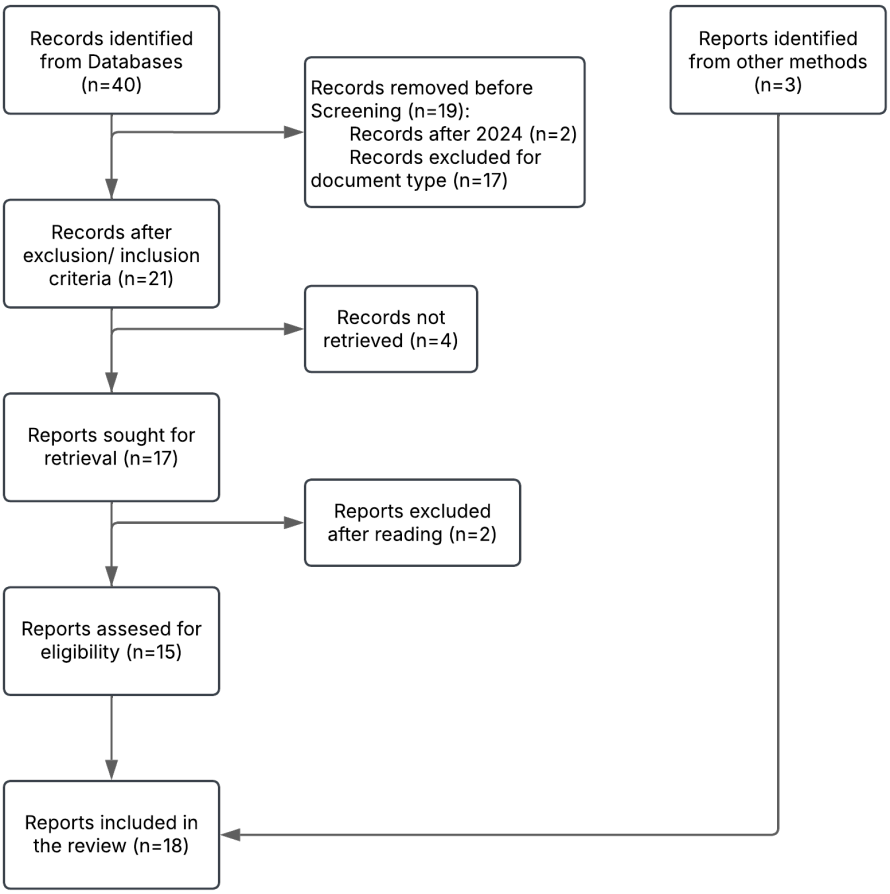


FIGURE 2.3. PRISMA 2020 Flowchart

two were in English, one because it was taken from a global commercial Massive Open Online Course (MOOC) platform where usually the reviews are in English [4] and the other one because it dealt with geo-located Tweets from United States of America [5] and one in Chinese, since it is geographically focused in China [6].

Considering that almost everything available related to NLP is in English, it is crucial to focus on the documents that did not use an English corpus to explore what has been done and what kind of tools were used to overcome that obstacle. The document that uses Portuguese corpora used Python's library SpaCy, which has multilingual/ Portuguese resources such as NER (Named-Entity Recognition) and lemmatization lists [7]. Although three documents use Chinese corpora, only one briefly addresses how it got around the issue. It is used BERTopic, this technique will be explained in more detail in Section 2.3.3, specifically the People's Daily pre-trained model, this model was trained using People's Daily, a Chinese newspaper so it understands better the formal Chinese language. At last, the model was fine-tuned for policy texts [8]. Contemplating now the documents that deal with multilingual textual data, this means that the corpora consist of texts in multiple languages, the simple way to overcome this was to translate the non-English text to English using an R package [9]. However, this is not always ideal because

some information might be lost or poorly translated because of characteristics already mentioned. In addition, the other two documents with multilingual textual data chose the same approach, using BERTopic selected multilingual sentence transformer [10], [11]. Overall, even though there is just one example with data in Portuguese, there are some solutions that can be applied in multiple languages, such as BERTopic.

2.3.2. Topic Modeling for Suggestion Texts

Textual data collected through questionnaires or open-ended responses, where respondents provide answers in their own words without predefined options [12], exhibit distinct characteristics compared to structured texts such as articles or newspapers. One of the differences is the need for context, in this case of the question, to understand the response as intended, whereas in articles or newspapers, the context is usually given in the full text. Furthermore, it is clear survey answers require a lower level of formality than articles. Additionally, questions performed during questionnaires can be non-factual, which requires an opinionated response [13]. Lastly, the answer written down by the respondent might be formulated in perfect sentences, but it may also consist of catchwords and contain grammatical and spelling mistakes [12]. This is also stated by other authors who declared that grammatical errors can cause noise that impairs text analysis and possibly be a source of incorrect or even false lemmas [7]. A possible solution is training models with noisy documents to increase performance and achieve better results [14]. All these attributes are essential to consider when working with this type of textual data, as they help address potential issues that may arise.

Despite the differences between the two types of textual data stated, the steps used to evaluate textual data obtained from suggestions are identical to the steps used to evaluate the other kinds of data. For instance, the basic treatment of textual data related to preprocessing steps as text normalization, conversion of all text to lowercase, and removal of stop-words, were all performed by the authors of [4], [5], [7], [15]–[18]. The choice between Stemming [4], [16], [17] and Lemmatization [7], [18] varied among authors, with some choosing not to apply either [5], [15].

Nevertheless, these steps are similar, and there are no significant differences worth relating. The main distinction comes from dealing with short-text data, which is going to be reviewed in the next Section.

2.3.2.1. *Topic Modeling in Short-Texts*

The length of the text to be analyzed has implications for the success of NLP tasks, in this case, topic modeling. Since the short-text has fewer words, it is possible that it may not contain enough meaningful words [19]. Short-texts lack word frequency and context information, causing severe sparsity problems for conventional topic models [20].

This problem is claimed by many authors who overcome it in different ways. Some decided to avoid this problem, for example, removing documents with less than a minimum words [4], [5], the author of [5] decided to remove documents with less than four words,

and the author of [4] removed documents with less than 40 words. There was a second approach to this question, by using models more appropriate to this type of text, such as the Biterm Topic Model, which directly models the word co-occurrence patterns based on biterms [20], used by the authors of [6]. ST-LDA, Latent Dirichlet Allocation of single topics, was also used, which is a variant of the topic modeling technique LDA [21] that assigns only one topic to the document with a membership probability [5]. Lastly, BERTopic is an embedding-based model, it focuses on capturing the semantic relationship in a lower-dimensional vector space [22], so it is understandable that it would perform well in short-text documents where the co-occurrence is minor. So, it is used by the authors of [10], [11], [18].

2.3.3. Topic Modeling Methods

Upon reviewing the documents, three predominant models were applied to perform topic modeling: LDA, Structural Topic Modeling (STM), and BERTopic. In Figure 2.4, it is possible to observe the evolution of the methods used by year in the final documents analyzed. In addition, it is possible to observe that BERTopic has gained importance in recent years.

LDA is a generative probabilistic model of a corpus. It assumes that each document is represented as a mixture of latent topics, and each topic is distinguished by a distribution over words [23]. LDA was the most used method, and it was applied in eight of the documents explored. This model was chosen due to its simplicity and its application in numerous fields [15], [24]. Another reason for its choice is the improvements compared to Latent Semantic Analysis (LSA) as they are the most known methods [7]. The necessity of choosing the number of topics is a drawback to this method. There are multiple strategies, including using Perplexity to evaluate the coherence of topics where a lower value denotes a better probabilistic model [25]. This approach requires a more exploratory analysis, considering a range of possible K values (number of topics) [9], [26]. There were other ways to explore the Perplexity, for example, the authors of [15] conducted a ten-fold cross-validation to determine a range of K values. Another approach involves selecting the number of topics through iterative experimentation and manual evaluation of topic interpretability [17], [24]. Some authors used Cosine-Similarity, which clusters by similarity all the topics generated and identifies the stable number of K [9]. Others used a coherence score that, similarly to Perplexity, implies the need to conduct an exploratory analysis [27]. Finally, the authors of [7] opted to use Grid Search to find the number of topics with the highest log-likelihood and the lower Perplexity.

Two documents used LDA adaptations: Dynamic Topic Model (DTM) and Single-Topic Latent Dirichlet (ST-LDA). DTM can reveal the topic evolution details of a large and unstructured document [28]. ST-LDA assigns to each document only one topic, with a membership probability which means the higher the more clear that topic is in that document. To evaluate the coherence of topics, the Point-wise Mutual Information score (PMI) was utilized by the authors of [5].

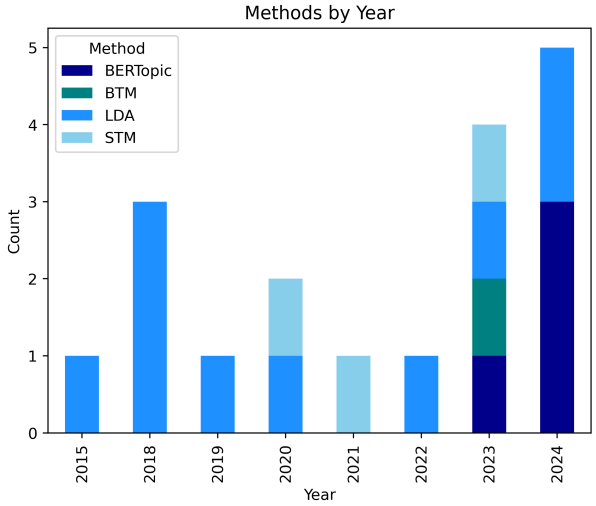


FIGURE 2.4. Topic Modeling Method used by Year

STM is used to perform text analysis and combines document metadata with topic modeling. This method mirrors the approach used by social scientists, who analyze various types of data to uncover relationships between variables [29]. STM was applied to account for the extracted topics and temporal variations they may have undergone [30], [31]. This method was chosen because it had already been introduced in political science [16].

BERTopic facilitates document embedding extraction using a sentence-transformer model that supports over 50 languages [32]. The authors of [10], [11], as mentioned in Section 2.3.1, used Multilingual BERTopic to process textual data in different languages. BERTopic is preferred over traditional topic modeling methods due to its ability to handle data scarcity and identify semantic connections between words, which is particularly beneficial for short-text documents [18]. Additionally, its flexibility in not requiring a predefined number of topics provides an advantage over LDA [33]. It is also possible to fine-tune parameters to achieve a higher topic coherence score, as demonstrated by the authors of [8], where variables like n_neighbors and min_cluster_size were adjusted. *GPT-3.5* was also integrated to generate comprehensive semantic topics related to the identified ones [8].

Looking at the methods used to evaluate the quality of the topics generated and the quality of each experiment, some authors revealed multiple quantitative approaches, but also included human verification and a quality evaluation [10], [11], [24]. The quantitative evaluation methods included Perplexity [7], [18], [26], [28] and Coherence score [6], [10], [11]. In some cases, authors decided to rely exclusively on human annotators and manual verification to evaluate the quality of the experiments [4], [5], [9], [18], [31].

2.4. Synthesis

There is a lot to bear in mind when dealing with topic modeling, such as the language, source, and length of the data. In this review, it is possible to understand how to deal with certain particularities of the data that is going to be processed, such as how to

deal with non-English text: by translating it or using specialized models like BERTopic that are available in multiple languages. Unlike other textual data, survey data requires contextual understanding, handling informal language, and addressing spelling errors. Even with these challenges, the preprocessing steps are similar between these kinds of textual data. Survey responses are usually concise, which means they might lack word frequency, which is challenging for traditional topic modeling methods like LDA. To overcome this obstacle, authors opted between removing short documents, using specific models like the Biterm Topic Model and ST-LDA, or applying embedding-based approaches like BERTopic. Lastly, models were compared, and even though LDA is the most used topic modeling technique, it has some drawbacks when compared with BERTopic, such as the difficulty of dealing with data scarcity and the need to choose the number of topics to generate. The evaluation of the topics is also a key aspect to decide which experiment yields the best result, it becomes clear that although there are some metrics available to evaluate the quality of topics, authors do not discard human evaluation.

Answering the LRQ1 research question, as stated in Section 2.2, there is a noticeable growth in topic modeling. There are also three major areas: topic modeling methods such as information extraction or text summarization, techniques used in NLP, for example Bidirectional Encoder Representations from Transformers (BERT) [34] and word2vec, and practical applications, for instance, social media and Covid-19. Responding to LRQ2 and LRQ3, Topic Modeling applied to Public Policies has a variety of uses, from supporting decision-making to comprehending the concerns of the citizens through social media. Even though this is a relatively explored area, only one document handled data in Portuguese, showing a gap in studies within this field and language. Although an exploration of short suggestions about public policies in Portuguese has not been done, there are comparable things, such as dealing with texts in Portuguese or handling short-texts that can be incorporated with each other to achieve this goal.

[This page is intentionally left blank.]

CHAPTER 3

Development

In this chapter, the work continues to follow the CRISP-DM methodology. Section 3.1 focuses on Data Understanding, where an analysis of the data will be made, as well as its characterization and its origin. Section 3.2 addresses Data Preparation, the data are processed for modeling based on the insights gained from the previous section. In Section 3.3 Modeling is covered, which involves selecting the most suitable models and trying to determine the optimal number of topics. Finally, Section 3.4 presents the Qualitative Evaluation, where the quality of the experiments is assessed by examining the resulting topics and their top word.

3.1. Data Understanding

The data used in the experiment was briefly explained in Section 1.1, and are retrieved from multiple workshops carried out by PLANAPP. During the workshops, participants, researchers, science communicators, science managers, and doctoral candidates affiliated with various research centers nationwide, were asked to write down answers to multiple questions that are next described. Each question led to an answer that was later transcribed to an excel workbook.

(i) "What challenges prevent scientific contributions from being acknowledged?"
(ii) "Do the identified challenges pertain to science, to public policy, or to both?"
(iii) "What solutions can be proposed for the identified challenge?"
(iv) "What strategies can be adopted for impactful communication?"
(v) "What concrete actions can be implemented to bring researchers closer to public policy makers?"

Questions (i) and (ii) produced three worksheets ‘Science’, ‘Public Policies’ and ‘Both’ which represented the challenges identified within these domains, questions (iii), (iv) and (v) originated a worksheet each: ‘Solutions’, ‘Strategies’, and ‘Actions’.

Table 3.1 presents a characterization of each worksheet after removing null / NA values, with the number of documents (records) and the average words per document. In addition to the answers to each question, each worksheet also includes the workshop ID, the research center ID, the year the workshop took place, the district where it was held, and the scientific area of the respondent. However, these data were discarded, as they were not relevant to the objectives of this study.

The documents are primarily in Portuguese, although some records are exclusively in English. Moreover, several Portuguese texts use technical vocabulary in English. This is something that should be taken into account in the following phases, especially in Data

TABLE 3.1. Data Characterization

Worksheet	Number of Records	Word Average per Record
Science	224	5.54
Public Policies	285	5.24
Both	259	4.89
Solutions	424	6.27
Strategies	379	5.75
Actions	239	7.48

Preparation (Section 3.2). Since the data were written on sticky notes in a workshop, although being text, the phrases are very short and informal in nature, containing abbreviations, for example, "inv" meaning "investigação" or "investigation" and symbols such as "+" instead of the word "mais", or "and" / "plus". Therefore, being usually shorter than short-text, we decided to call it, over this document, "tiny text".

Figure 3.1 shows a word cloud for each worksheet without any preprocessing. By analyzing the word clouds, it becomes clear that the removal of stop words will be a crucial step in Section 3.2 for the preparation of the data, as the most prominent words are stop words, with particular emphasis in "de" which is a discourse connector frequently used in Portuguese. Moreover, it is evident that, although each word cloud represents the responses to a different question, several words appear repeatedly across the figures, which justifies combining them during the data preparation stage.

(A) Science (B) Policies (C) Both

(D) Solutions (E) Strategies (F) Actions

FIGURE 3.1. Word Clouds of the Unprocessed Data by Worksheet

3.2. Data Preparation

Since there is not much data available for each category and the themes seem to recur across the different categories, all records were joined into one corpus with 1810 documents and an average of 6.43 words each. Figure 3.2 displays word clouds from this corpus using unigrams, bigrams and trigrams.

Although the stop words have not been removed yet, there is already some valuable information to take from the word clouds represented in Figures 3.2b 3.2c. The most common bigram is "falta de" ("lack of"), appearing a total of 98 times. This expression conveys the definition of the problems that would be subsequently referred. The most common trigram appears 12 times, being "canais de comunicação" ("communication channels"), which points to both a potential solution and a recurring issue related to communication between science and public policy.

(A) All Data Word Cloud

(B) Bigram Word Cloud

(C) Trigram Word Cloud

FIGURE 3.2. N-grams Word Clouds

A challenge found in Section 3.1, was the existence of abbreviations. To try to solve this problem, a dictionary of abbreviations was created. The dictionary has the contraction as the key and its meaning as the value, e.g., "ong": "organização não governamental".

Another issue that emerged during the data understanding stage was that some of the documents were written in English. To solve it, the *translators Python library* [1] was used, leveraging *Google Translate* as the underlying translation engine. The process was to first detect the language of the document and, if English was detected, it would be translated to Portuguese. However, if a text was predominantly in Portuguese but contained a few words in English, these words would not be detected and the text would remain unchanged.

To prepare the data for modeling, six datasets were generated. The first and most simple is called "No Preprocessing" where the only step was to tokenize the documents, without any additional step involved. The second, "Simple Preprocessing", implemented

[1] *translators* version 6.0.1, PyPI, https://pypi.org/project/translators/

the abbreviation dictionary, transformed all words to lowercase, removed digits and punctuation. In the third dataset, named "Normal Preprocessing", besides the preprocessing steps done for "Simple Preprocessing", all stop words were removed. The fourth dataset is a variant of the third but the documents in English are translated to Portuguese and it is called "Portuguese Normal Preprocessing". The fifth, "Total Preprocessing" in addition to the steps involved in "Normal Preprocessing", also preforms lemmatization. In line with "Normal Preprocessing" and "Portuguese Normal Preprocessing", there is also a sixth dataset which is the translated version of the original dataset. In the last two datasets, lemmatization was chosen over stemming, as stemming often produces excessively shortened words that can become unintelligible. Moreover, given that one of the methods applied in the modeling stage is BERTopic, which relies on embeddings rather than word frequency for topic modeling, lemmatization was deemed more appropriate. Table 3.2 represents the preprocessing steps in a simple way.

Table 3.2. Datasets Preprocessing Steps

Dataset	**Preprocessing Steps**
No Preprocessing	Tokenization
Simple	No Preprocessing + Abbreviation Dictionary + Lowercase + Removal of digits and punctuation
Normal	Simple + Removal of Stop Words
Portuguese Normal	Normal + Translation to Portuguese
Total	Normal + Lemmatization
Portuguese Total	Total + Translation to Portuguese

Table 3.3 presents the average number of words per document in each dataset. These values reveal important observations, such as the expected increase in word count when applying the abbreviation dictionary. Furthermore, translation into Portuguese also leads to a higher average word count per document, since Portuguese makes more frequent use of articles and discourse connectors compared to English.

Table 3.3. Datasets Word Average per document

Dataset	**Average**
No Preprocessing	5.88
Simple	5.89
Normal	4.08
Portuguese Normal	4.12
Total	4.08
Portuguese Total	4.13

3.3. Modeling

This section details the modeling stage, starting with the determination of the ideal number of clusters or, in this case, topics, in Section 3.3.1, followed by a description of the topic modeling techniques applied in Section 3.3.2.

3.3.1. Ideal Number of Topics

The first step of modeling is to decide the number of topics the models should generate. The most used approximation technique is to use the Elbow method [35] in combination with K-means [36], to generate a curve where, as the value of K (number of clusters) increases, the curve decreases until reaches a stable point [37]. The value of K at this point is the ideal number of clusters. Another way to try to identify the ideal number of topics is to use the Silhouette method [38] which combines separation and cohesion, and the higher the Silhouette score the better. As described in [39], the Silhouette values range between -1 and 1. Both these methods were applied in order to attempt to identify the optimal number of clusters for the experiments. To perform this task, the "Simple Preprocessing" dataset was used, and both, Term Frequency-Inverse Document Frequency (TF-IDF) [40] and Bag-of-Words (BoW) Vectorizers [41] were employed. The K values were sought in the range $\{3, \ldots, 35\}$. The results obtained through these experiences are shown in Figures 3.3, 3.4.

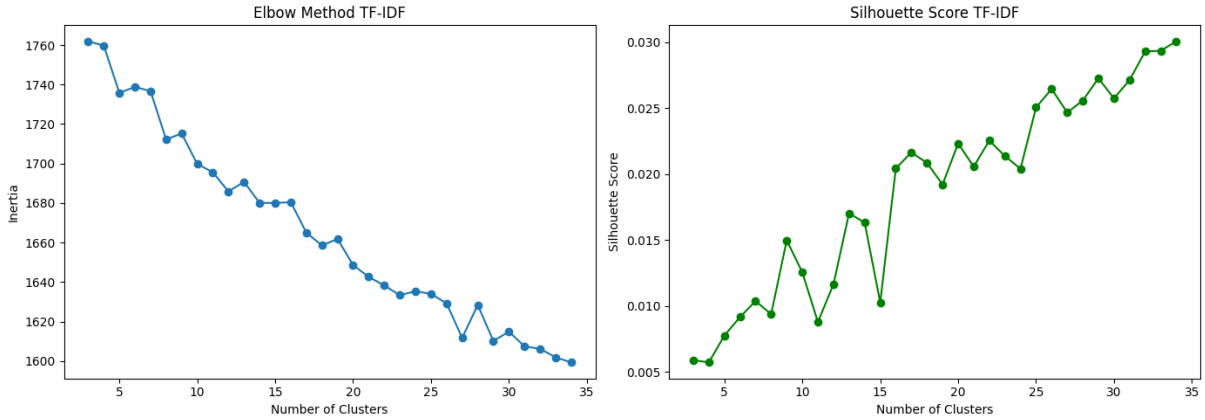


FIGURE 3.3. Elbow curve and Silhouette Score using TF-IDF

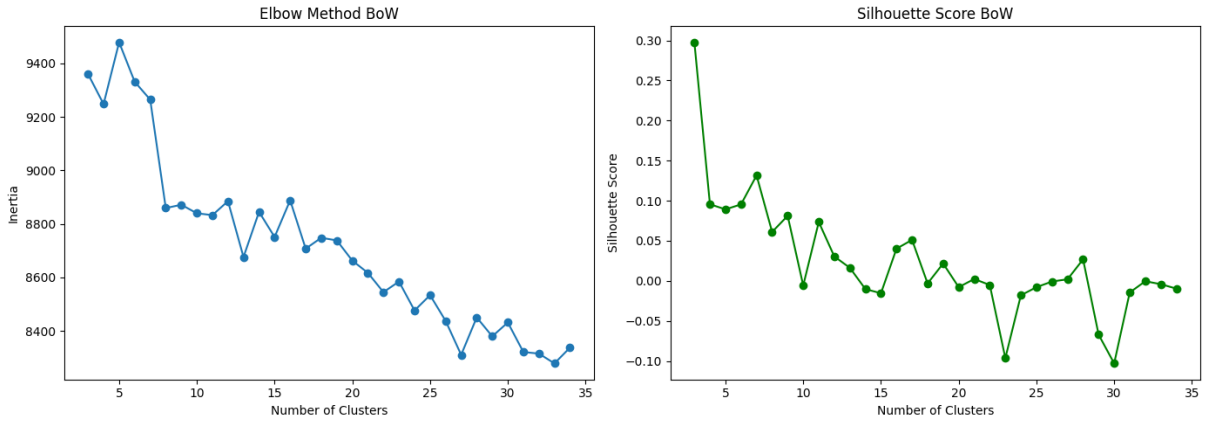


FIGURE 3.4. Elbow curve and Silhouette Score using BoW

Analyzing the graphs and trying to apply the empirical rule of choosing the point where some plateau begins for the Elbow method or the highest score for the Silhouette

technique, it is not possible to draw any conclusions about the ideal number of clusters. The graphs obtained through the Elbow Method did not reveal the expected point for an "elbow", and the graphs for the Silhouette method had a maximum value around 0.03 using TF-IDF and 0.3 using BoW, which is too close to 0 to be able to consider it a good result.

Without being able to determine the ideal number of clusters, it was decided to employ four different values: 5, 10, 15, and 20 clusters. The goal is to make enough experiments to be able to properly evaluate the topic models and finally choose the one with the best performance.

3.3.2. Models

The models chosen to apply to the data, based on the Literature Review, Chapter 2, were LDA, even with challenges such as, the scarcity of the data or the lack of length of the documents, since it is a classical method and a solid starting point for experimentation.

Beyond LDA, BERTopic was also selected as it appears to be more suitable for the data under the study, being embedding-based it does not need as much data as LDA. However, challenges may still arise because the documents consist of tiny text and consequently may lack sufficient context, which is particularly important for BERTopic.

For LDA, the BoW vectorizer was used in order to keep this experiment as classical and simple as possible. For BERTopic, two different sentence-transformers were used Multilingual, it was chosen since it has more than 50 languages available [32] and the majority of documents are in Portuguese, this sentence transformer is particularly useful to apply to the datasets that are not translated. The second sentence-transformer was AlBERTina which is totally pretrained in Portuguese text. It was chosen to check if there would be differences between the Multilingual and a more Portuguese focused sentence-transformer. Although BERTopic does not have the need to choose the number of topics, to have a fair comparison between models, the same number of topics were generated by using K-means.

In total, 72 experiments were conducted by combining the six datasets, the four different topic settings tested (5, 10, 15, and 20 topics), and the three model variations.

3.4. Qualitative Evaluation

When trying to evaluate the topics qualitatively, by observing the top ten words of each topic, it is possible to note that the experiments with the "No Preprocessing" dataset had more stop words, making it harder to understand what the themes of the topics were. In addition, when using the referred dataset and when performing the topic modeling with LDA, the top 10 words had a lot more noise, including special characters such as "," and "-". Furthermore, when using AlBERTina sentence-transformer and the datasets that do not translate the text, there is usually a topic which consists in only English words, mainly connectors. Another relevant point to mention is the experiments with 5 and 10

topics seem to have topics more focused and without overlapping themes between each other.

Although some valuable insights can be retrieved from observing the topics, it is impossible to compare all 72 experiments and draw a clear conclusion about which one performed best. Therefore, Chapter 4 explores various methods for quantitatively evaluating the experiments.

[This page is intentionally left blank.]

CHAPTER 4

Quantitative Evaluation

This Chapter focuses on determining which combination of datasets, number of topics, and model variation produces the best results quantitatively. However, in the context of topic modeling, what constitutes the "best result" is not always clearly defined. These conclusions can be achieved through the use of some metrics such as Coherence and Perplexity, or alternatively through human evaluation, using annotators, to assess whether the topics are coherent and well separated. Nevertheless, it is understood that a well-defined set of topics implies that the documents within each topic address closely related themes, while documents represented by different topics are distant from one another.

4.1. Coherence Score

Figures 4.1 and 4.2 show the evolution of the coherence, using CV and UMass respectively, considering each combination of the preprocessing dataset with each number of topics. The graphs were separated between sentence transformer, where the default model is the multilingual and AlBERTina model is the Portuguese, for a better readability and to try to understand which has higher values of coherence.

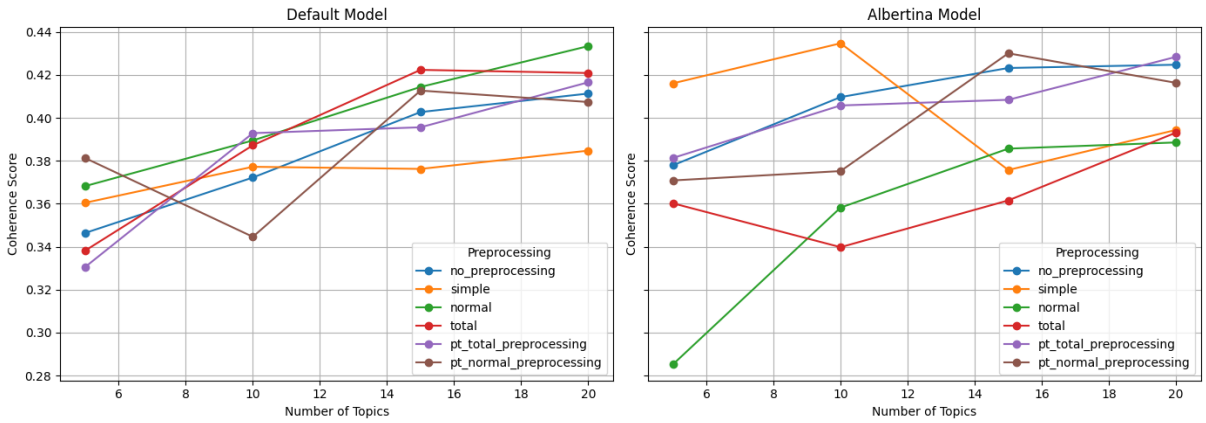


FIGURE 4.1. BERTopic Coherence (CV) Results

Some notes should be taken from the visualization of the graphs in Figure 4.1. The first one is that none of the 48 combinations (2 models x 6 datasets x 4 topic settings) had values of coherence score that are considered good, usually close to 1 since it varies between 0 and 1. The second observation is that, when using this metric to compare the two models, the Multilingual model generally performs better than the Portuguese one, although the difference is not substantial. Furthermore, the translated datasets reveal significant improvements when using the AlBERTina sentence-transformer, unlike Multilingual, where the non-translated datasets have a slightly better performance. Lastly,

there is no preprocessing dataset that seems to have better results, since the better preprocessing step varies between the models used. When using AlBERTina the best preforming preprocessing varies between topic settings being simple, with 5, 10 topics, normal in Portuguese with 15 topics, total in Portuguese with 20 topics.

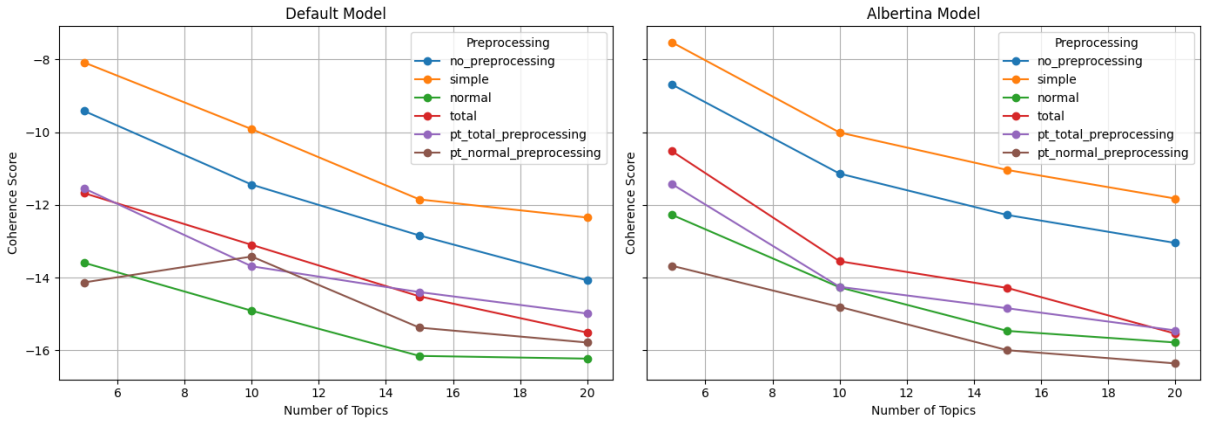
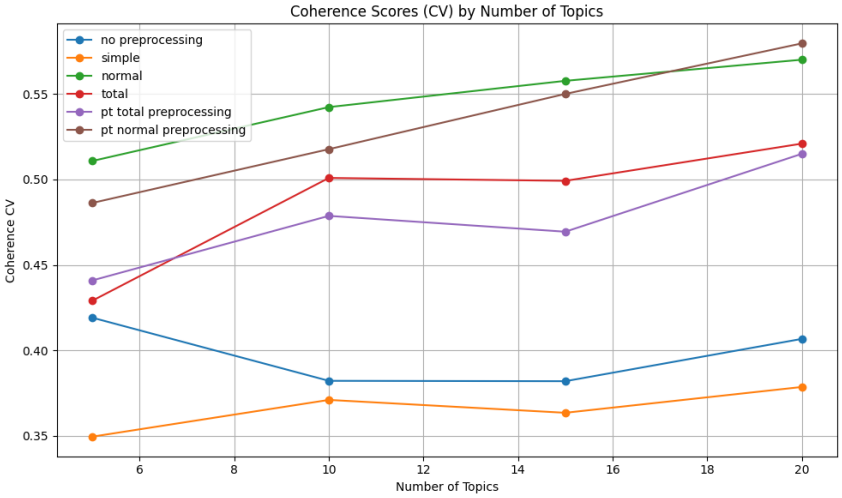


FIGURE 4.2. BERTopic Coherence (UMass) Results

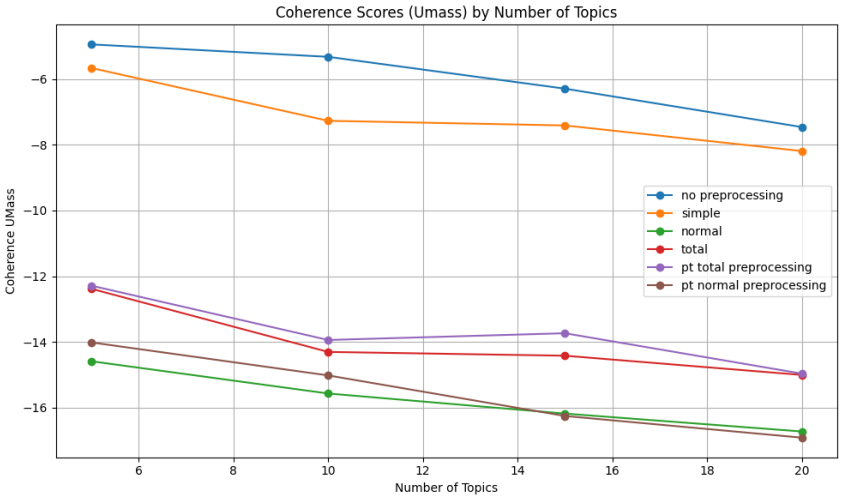
In Figure 4.2 Coherence was calculated by using UMass which the ideal value is 0. These results cannot be considered good because the highest values are slightly over -8 which is still not close enough to the ideal value, 0. Comparing the models and the preprocessing datasets it is clear, by evaluating by this metric, the simplest preprocessing datasets, Simple and No preprocessing, are the ones preforming the best. The best preforming topic setting is using only 5 topics. However, this is expected as the coherence values decrease while the K (number of topics) increases [42]. The translated texts do not have an impact when compared to the originals, and interestingly when using multilingual sentence transformer, the Portuguese normal dataset has better results than the original contrary to when using AlBERTina where the non-translated actually has better results. Using UMass, Simple and No preprocessing datasets stand out from the rest mainly using AlBERTina sentence-transformer.

Figure 4.3a shows that in this experiment, Normal preprocessing, original and translated, presents better results than the other datasets. It is also noticeable that the values are higher using LDA instead of BERTopic in contrast to what would be expected given the reduced size of the dataset. Moreover, coherence increases as the number of topics increases, which may be questionable considering the limited number of documents. This might mean that instead of trying to find larger themes, the model is splitting the documents into overly specific topics. In doing so, although coherence increases, human interpretability tends to decrease, particularly when working with such small datasets.

The results shown in Figure 4.3b are similar to the results of Figure 4.2, where the simpler datasets perform noticeably better than the others. The coherence values are slightly higher than the ones using BERTopic (close to -5), which is an improvement, although it is still not enough to conclude that these are the best experiments.



(A) LDA Coherence (CV) results



(B) LDA Coherence (UMass) results

FIGURE 4.3. LDA Coherence results

Overall, it is not possible to identify a combination, or even individual preprocessing strategy, topic configuration, or model, that consistently produces the best results across all scenarios. Coherence metrics point out the models struggle to achieve values that can be considered satisfactory, which reflects the challenges of topic modeling on short and sparse texts. While the Multilingual sentence-transformer generally performs slightly better than the Portuguese model, the differences are not substantial, the translated datasets present some advantage over the originals, although they are not clear in all tests. LDA also has its problems because the human interpretability might decrease, while the coherence increases. This reinforces that coherence alone is not a reliable metric to assess topic quality in this context.

4.2. Perplexity

Perplexity is another way to evaluate the quality of topics generated in a given experiment, since it measures the level of confidence of the model in its predictions. This metric

is commonly used to evaluate classical models but is not well defined for masked language models such as BERT. Therefore, it was computed only for the LDA experiments. Perplexity values are interpreted where lower scores indicate better performance and a more generalized model.

Table 4.1 shows the results of perplexity for each LDA experiment, showing that none of the experiments performed well in this metric. Since the lowest value, 773.41 with Simple preprocessing and 15 topics, remains considerably high. Furthermore, the tendency for lower values to occur with 20 topics may indicate overfitting, similarly to what was observed with Coherence.

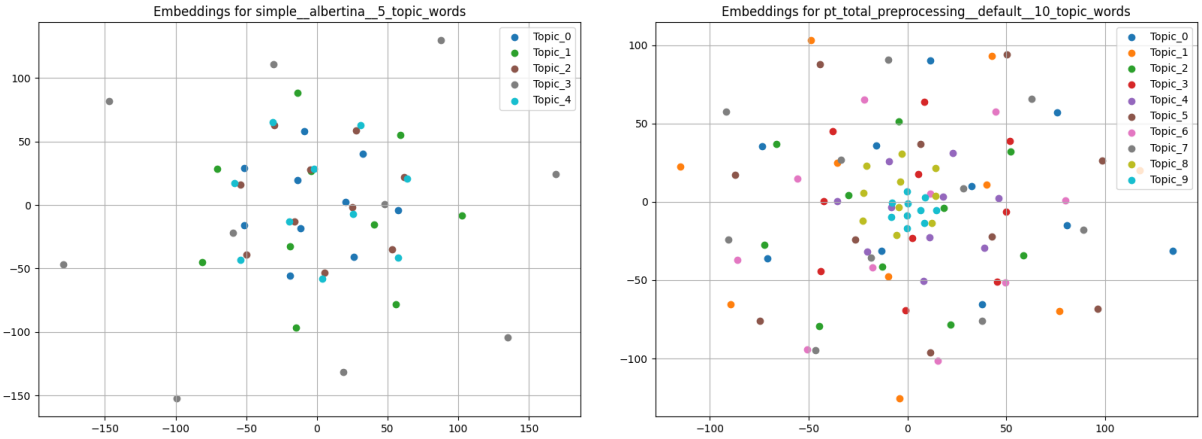
TABLE 4.1. LDA Perplexity results

Preprocessing	Number of Topics	Perplexity
No Preprocessing	5	933.61
	10	924.20
	15	883.93
	20	857.29
Simple Preprocessing	5	835.69
	10	816.41
	15	773.41
	20	777.30
Normal Preprocessing	5	1518.00
	10	1489.84
	15	1393.07
	20	1352.42
Portuguese Normal Preprocessing	5	1399.81
	10	1399.50
	15	1327.45
	20	1290.81
Total Preprocessing	5	1073.02
	10	1087.42
	15	1049.13
	20	1015.87
Portuguese Total Preprocessing	5	1064.54
	10	1051.00
	15	1030.45
	20	973.82

4.3. Visualizing Topics

After using the most classical evaluation metrics and the difficulty of retrieving any conclusions on which experience yields the best results, it was decided to try to visualize the topics as embeddings. This is, having the top 10 representative words of each topic, the embeddings were calculated using "marquesafonso/albertina-sts" sentence-transformer, then it was applied dimension reduction to project the data in two dimensions. The purpose was to visualize the embeddings and try to evaluate the topics from that visualization. This approach has problems, such as, the embedding of certain words can change

depending on the context of the sentence where they appear, and here we are taking them out of that context. Another problem of this approach would be the challenge of reducing the dimensionality of the embeddings, where there is always a loss of information. However, this was still tested and two examples are shown in Figure 4.4.



(A) Simple Preprocessing with AlBERTina and 5 topics

(B) Portuguese Total Preprocessing with Multilingual and 10 topics

FIGURE 4.4. Embedding Visualization

The visualization of the figures from this experiment suggest that this method is not effective for topic evaluation, as there is no clear separation between topics or cohesion within each topic, with the exception of topic 9 in Figure 4.4b where all points are compacted in the same area.

4.4. Shannon Entropy

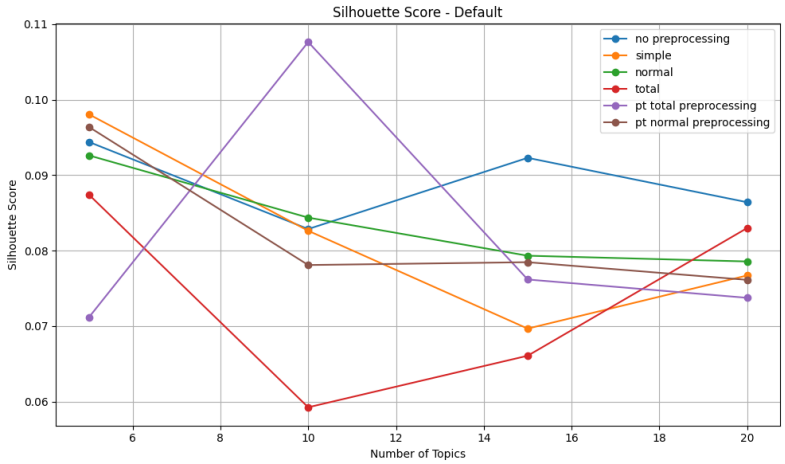
The Shannon entropy was then used to evaluate the variance of a topic within an experiment. Higher entropy indicates more diffuse topics that are harder to interpret, whereas lower entropy reflects more specific and interpretable topics. In this case, the entropy will vary between 0 and $\log_2(x)$, with x the number of top words analyzed, in this case 10. However, all results were close to the maximum entropy, which is not desirable.

4.5. Silhouette Score

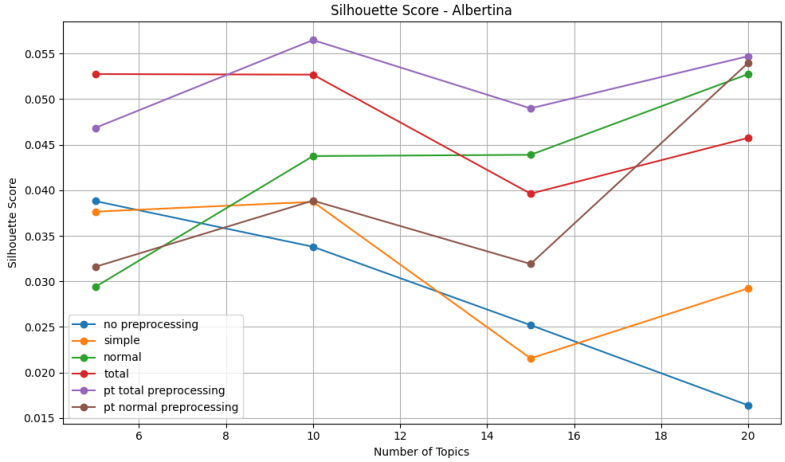
The silhouette score [43] is used to assess how well clusters are separated. This metric is not usually used to evaluate topic modeling, because even if the clusters (topics) are well separated, the topics can be badly formed and not semantically related. However, after trying the classic methods for evaluating topic modeling and not having success, we tried to evaluate it with this clustering evaluation technique. The cosine distance was the metric chosen to calculate silhouette score since the points are represented by vectors and their angle play a bigger role than their length.

For BERTopic, the silhouette score was calculated using the embedding values of the documents. Given the high dimensionality of embeddings, it is expected the resulting scores are relatively low. Figures 4.5b and 4.5a present the silhouette score results, with

the multilingual model with slightly higher results. However, as expected, generally the results are quite low.



(A) BERTopic Multilingual Silhouette Score



(B) BERTopic AlBERTina Silhouette Score

FIGURE 4.5. BERTopic Silhouette Score by Sentence Transformer

Since LDA attributes probabilities of belonging to each topic, this means, it does not attribute a single topic to a document. To compute silhouette score we need to attribute each document to a single topic. To do this, the topic attributed to each document is the most probable topic for each topic. So, documents from the same topic have nearly identical probability vectors, on the other hand, documents from different topics are almost orthogonal. Using cosine-similarity, this implies that the silhouette score will be very close to 1 - which is the maximum value. In Figure 4.6 it is possible to see the values of the silhouette score by preprocessing, and number of topics. As previously referred, the values are close to 1 which is questionable, especially considering the results of other metrics.

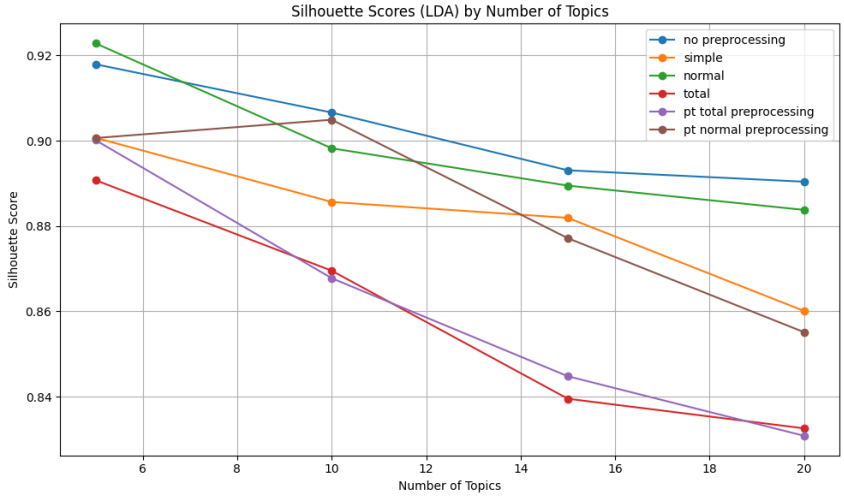


FIGURE 4.6. Silhouette Score for LDA

4.6. Using Cosine-Similarity

Another way that was tested with the goal of evaluating the quality of the topics of each experiment was by using cosine similarity with embeddings to evaluate both: the similarity within the topic and the similarity between topics.

When evaluating the similarity within the topic, higher values are desirable, as they indicate that the most relevant words are closely related. The process was to calculate the cosine similarity between each pair of words within the same topic, and the average value was used as the overall measure. This returns a value for each topic. The minimum and maximum values of cosine similarity were also saved for comparison since the mean can be too penalizing. Figure 4.7 illustrates a snip of the heatmap with these results when using the BERT. The full version of the heatmap is provided in Figure A.2 for reference. The graph shows that even though there are some topics with high cosine similarity, most topics in each experiment have poor results.

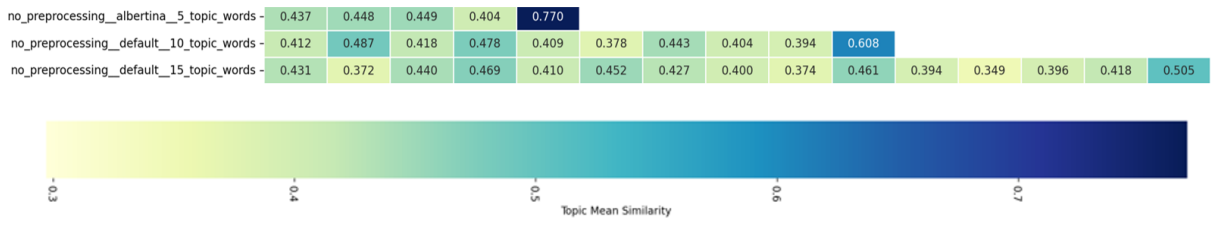


FIGURE 4.7. Excerpt of the Cosine Similarity Heatmap

Contrary to the similarity within topic, when evaluating the similarity between topics, the goal is to be as low as possible, since that would mean that the topics are well separated from each other. To compute this, the mean of the embeddings of the top ten words of each topic was taken, and now, assuming each mean embedding as a topic, the cosine similarity between each topic was calculated, after this, for each experiment it was determined the mean, the minimum and the maximum.

This approach provides an oversimplified measure of the difference between topics, primarily because averaging the embeddings consolidates all information into a single vector, which can hide much of the original meaning. The results of this experiment were higher than desired, ranging between 0.7 and 0.9, indicating poor topic separation, although this result was expected considering the problems of this calculation.

It is important to note that, albeit none of the attempts to evaluate the quality of the topics generated worked or achieved ideal results, this does not imply that the models produced 'bad' topics or that they were unable to identify any topics. Instead, the metrics used to evaluate those topics are not suited for evaluating the kind of data in this experiment: tiny text and very limited records.

When evaluating tiny text and the kind of problems we have been facing, other authors rely on human annotators to evaluate and define the best experiments. However, it would be interesting if we could evaluate these challenging corpus quantitatively. With this objective in mind, we have developed the measurement *Singularity Score (SS)* described next (Section 4.7).

4.7. Singularity Score

Singularity Score arises to fill the existing challenges when evaluating tiny text such as the ones we encounter in this work. The main goal of the SS is to emulate the behavior of human annotators, since the ultimate objective of a topic modeling experiment is to be the most interpretable to humans as possible.

The premise was to use the top N – we consider 10 to be the ideal value but it may vary – significant words from each topic as a starting point. We consider a good set of topics, the one in which the themes of one topic do not appear in the others, and each topic is relatively specific to a subject.

Using the top N significant words of each topic, returned by the model, we count the number of unique and non-unique words in a topic compared to the other topics in the model and combine it with the fact that the topic has or does not have one or more significant words. As the top words retrieved by a model are not repeated, we consider the stem of those words. This means that our evaluation will focus on the words being in the same family.

Let a topic model have N topics, and for each topic i:

- Significant Word, $SW \in \{0, 1\}$
- Count of Unique Words, $UW \in [0, 10]$
- Count of Non-Unique Words, $NUW \in [0, 10]$

Having a SW means that there is at least one stem repeated in the topic. UW is the number of stems repeated in the topic and do not appear in any other topic. Lastly, NUW is the number of stem repeated in the topic but appear in another topic.

Topic uniqueness (tu_i) is measured for each topic, varies between 0 and 1 and evaluates how unique a topic is. It rewards when the topic has significant words, has unique words, and has low non-unique words. Topic uniqueness is given by Equation 4.1.

$$tu_i = w_{SW} SW_i + w_{UW} \frac{UW_i}{10} + w_{NUW} \left(1 - \frac{NUW_i}{10}\right) \tag{4.1}$$

Where $w_{SW} + w_{UW} + w_{NUW} = 1$

The w of each metric can be fine-tuned to accommodate different problems, and reward one metric more than other depending on the context of the text data.

The topic uniqueness (TU) of the model can now be calculated by applying the average of the tu_i of each topic in the model, it is given by Equation 4.2.

$$TU = \frac{1}{N} \sum_{i=1}^{N} tu_i \tag{4.2}$$

However, a simple average might overly penalize a few weak topics, even when most topics have a high value of tu. In order to add weight to scenarios in which most topics demonstrate greater uniqueness, a piecewise function f was created. Let ST be the fraction of the Strong Topics, given by Equation 4.3, and θ the threshold of what it means to be "strong" (we suggest values around 0.7).

$$ST = \frac{number\ of\ topics\ with\ tu \geq \theta}{N} \tag{4.3}$$

Figure 4.8 is a general representation of the formula. When $ST > 0.5$ the SS increases linearly depending on the β value.

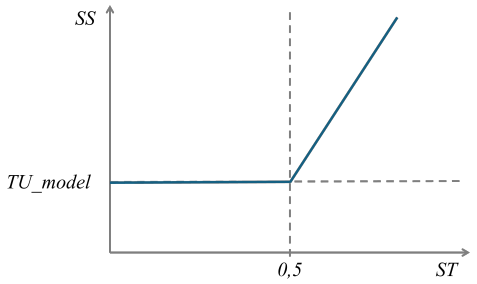


FIGURE 4.8. SS plot representation, TU_model is the TU value of the topic model, and the line represents the SS value depending on the ST (the fraction of the strong topics)

The function f applied to ST, demonstrated by Equation 4.4, assigns a higher weight to models in which more than half of the topics are strong (high values of tu). The β magnitude of this weighting factor ($\beta \in [0, 1]$).

$$f(ST) = \begin{cases} 0 & if\ ST \leq 0.5 \\ \beta ST & if\ ST > 0.5 \end{cases} \tag{4.4}$$

Finally, the SS is provided by the TU, and rewarded when most topics are strong, it is given by Equation 4.5.

$$SS = TU(1 + f(ST)) \tag{4.5}$$

4.7.1. Singularity Score Validation

To validate the SS, we decided to use topics from *Incorporating Word Correlation Knowledge into Topic Modeling* [44], where four annotators evaluate the coherence, the authors call it the Coherence Measure (CM) of the topics created. This measure represents the percentage of relevant words to a topic within the "candidate words" that are the ten top words from that topic. In this study, the authors are comparing the performance between models to decide which offers better results. This is relevant to our validation because it uses the top ten words and the annotators' validation to rank the models.

The topics used for the calculation of the CM for the 20-newsgroups dataset, as described in the reference [44] are presented in Figure 4.9 and the results (also as presented in [44]) in Figure 4.10.

LDA				DF-LDA			
Topic 1 (Crime)	Topic 2 (Sex)	Topic 3 (Sports)	Topic 4 (Health)	Topic 1 (Crime)	Topic 2 (Sex)	Topic 3 (Sports)	Topic 4 (Health)
gun	sex	team	government	gun	**book**	game	money
guns	men	game	money	police	men	games	pay
weapons	homosexuality	hockey	private	carry	**books**	players	insurance
control	homosexual	season	people	kill	homosexual	hockey	policy
firearms	gay	**will**	**will**	killed	homosexuality	baseball	tax
crime	sexual	**year**	health	weapon	**reference**	fan	companies
police	**com**	play	tax	cops	gay	league	**today**
com	homosexuals	nhl	care	warrant	**read**	played	plan
weapon	people	games	insurance	deaths	male	season	health
used	**cramer**	teams	program	control	homosexuals	ball	jobs
Quad-LDA				MRF-LDA			
Topic 1 (Crime)	Topic 2 (Sex)	Topic 3 (Sports)	Topic 4 (Health)	Topic 1 (Crime)	Topic 2 (Sex)	Topic 3 (Sports)	Topic 4 (Health)
gun	homosexuality	game	money	gun	men	game	care
guns	sex	team	insurance	guns	sex	team	insurance
crime	homosexual	play	**columbia**	weapons	women	hockey	private
police	sin	games	pay	child	homosexual	players	cost
weapons	marriage	hockey	health	police	homosexuality	play	health
firearms	**context**	season	tax	control	child	player	costs
criminal	people	**rom**	**year**	kill	ass	fans	company
criminals	sexual	**period**	private	deaths	sexual	teams	companies
people	gay	goal	care	death	gay	fan	tax
law	homosexuals	player	**write**	**people**	homosexuals	best	public

FIGURE 4.9. Topics of 20-newsgroups Dataset from *Incorporating Word Correlation Knowledge into Topic Modeling* [44]

Method	Annotator1	Annotator2	Annotator3	Annotator4	Mean	Standard Deviation
LDA	30	33	22	29	28.5	4.7
DF-LDA	35	41	35	27	36.8	2.9
Quad-LDA	32	36	33	26	31.8	4.2
MRF-LDA	**60**	**60**	**63**	**60**	**60.8**	**1.5**

FIGURE 4.10. CM (%) on 20-Newsgroups Dataset from *Incorporating Word Correlation Knowledge into Topic Modeling* [44]

The results presented by the authors in the reference [44] can be observed in figures 4.11 and 4.12. Namely, the topics used for the calculation of the CM for the NIPS dataset are presented in Figure 4.11 and the results in Figure 4.12.

LDA				DF-LDA			
Topic 1 (Vision)	Topic 2 (Neural Net)	Topic 3 (Speech)	Topic 4 (Circuits)	Topic 1 (Vision)	Topic 2 (Neural Net)	Topic 3 (Speech)	Topic 4 (Circuits)
image	network	hmm	chip	images	network	speech	analog
images	neural	**mlp**	analog	pixel	system	context	chip
pixel	feedforward	hidden	**weight**	view	connection	speaker	vlsi
vision	architecture	context	digital	recognition	application	frame	implement
segment	**research**	model	**neural**	face	artificial	continuous	digital
visual	**general**	recognition	hardware	ica	input	processing	hardware
scene	**applied**	probabilities	bit	vision	**obtained**	**number**	voltage
texture	**vol**	training	**neuron**	system	**department**	**dependent**	bit
contour	**paper**	markov	implement	**natural**	**fixed**	frames	transistor
edge	**introduction**	system	vlsi	faces	techniques	spectral	design
Quad-LDA				MRF-LDA			
Topic 1 (Vision)	Topic 2 (Neural Net)	Topic 3 (Speech)	Topic 4 (Circuits)	Topic 1 (Vision)	Topic 2 (Neural Net)	Topic 3 (Speech)	Topic 4 (Circuits)
image	training	speech	circuit	image	network	hmm	chip
images	set	hmm	analog	images	model	speech	**synapse**
pixel	network	speaker	chip	pixel	learning	acoustic	digital
region	learning	acoustic	voltage	disparity	function	context	analog
vision	net	phonetic	current	color	input	word	board
scene	number	vocabulary	vlsi	intensity	neural	phonetic	charge
surface	algorithm	phone	**neuron**	stereo	set	frames	**synaptic**
texture	**class**	utterance	gate	scene	algorithm	speaker	hardware
local	input	utterances	**input**	camera	system	phone	vlsi
contour	**examples**	frames	transistor	**detector**	data	vocabulary	programmable

FIGURE 4.11. Topics of 20-newsgroups Dataset from *Incorporating Word Correlation Knowledge into Topic Modeling* [44]

Method	Annotator1	Annotator2	Annotator3	Annotator4	Mean	Standard Deviation
LDA	75	74	74	69	73	2.7
DF-LDA	65	74	72	47	66	9.5
Quad-LDA	40	40	38	25	35.8	7.2
MRF-LDA	**86**	**85**	**87**	**84**	**85.8**	**1.0**

FIGURE 4.12. CM (%) on NIPS Dataset from *Incorporating Word Correlation Knowledge into Topic Modeling* [44]

To calculate the SS for both datasets, the exact same values of weight ($w_{SW} = 0.2; w_{UW} = 0.4; w_{NUW} = 0.4$), threshold ($\theta = 0.65$) and relative bonus ($\beta = 0.7$) were used, not only for these validations, but also for our topics, these results are available in Subsection 4.7.2. The results of the SS validation are displayed in Table 4.2, for the 20-newsgroups dataset and in Table 4.3, for the NIPS dataset.

TABLE 4.2. Singularity Score for 20-Newsgroups Dataset

Model	Singularity Score
LDA	0.99125
DF-LDA	0.96075
Quad-LDA	0.96075
MRF-LDA	1.258

TABLE 4.3. Singularity Score for NIPS Dataset

Model	Singularity Score
LDA	0.47
DF-LDA	0.54
Quad-LDA	0.54
MRF-LDA	0.47

The results obtained through the 20-newsgroups dataset, are able to validate SS since MRF-LDA is the model presenting the best score in both evaluations, CM (Figure 4.10) and SS (Table 4.2). Although the results of the other models are not ranked equally, this can be justified by the high values of standard deviation shown in Table 4.10. We are also able to observe in Figure 4.9 that within each topic in each model, there are different words with repeated stems, which is what SS relies on to evaluate each experiment.

When comparing the results of CM (Figure 4.12) and SS (Table 4.3) using the NIPS dataset, it is possible to notice that SS fell short. In particular, SS failed to capture the differences between models highlighted by the CM evaluation and was unable to identify the best-performing model. This result can be explained by the scarcity of words with the same stem, which, as previously mentioned, serves as the basis for SS computation.

The topics are considered good by the annotators because, although they lack words with the same stem, they have something equally as important to evaluate the quality of topics, semantic similarity among words.

Semantic similarity is not considered in the computation of SS, yet it can be as important as stem similarity. To address this limitation, we propose that the next iteration of the SS incorporates the calculation of semantic similarity using embeddings or even the usage of *PyDictionary* [1] library to check if any of the top ten words are synonyms.

4.7.2. Employ Singularity Score

To calculate the SS, the first step is to attribute the weight to retrieve tu. We have decided to give them the following values $w_{SW} = 0.2; w_{UW} = 0.4; w_{NUW} = 0.4$. Subsequently, to decide the threshold (θ), we tested with 0.55, 0.6, 0.65, and 0.7 and decided to go with $\theta = 0.65$ and the relative bonus $\beta = 0.7$. In Tables 4.4 and 4.5 we present the results of this metric.

Next, in Table 4.6 it is displayed the top 10 words of the best and worst preforming experiences evaluated with SS. The experiment with a higher SS is obtained by using

[1] `https://pypi.org/project/PyDictionary/`

TABLE 4.4. Singularity Score using LDA

Preprocessing	Number of Topics	Singularity Score
No Preprocessing	5	**0.4**
	10	0.424
	15	0.4187
	20	0.442
Simple Preprocessing	5	**0.4**
	10	0.464
	15	0.456
	20	0.426
Normal Preprocessing	5	0.512
	10	0.472
	15	0.4827
	20	0.446
Portuguese Normal Preprocessing	5	0.472
	10	0.508
	15	0.464
	20	0.458
Total Preprocessing	5	0.512
	10	0.424
	15	0.456
	20	0.438
Portuguese Total Preprocessing	5	0.52
	10	0.436
	15	0.4347
	20	0.44

the Portuguese dataset with normal preprocessing, BERTopic model, the Multilingual sentence-transformer and retrieving 10 topics, with SS of 0.942 (Table 4.5). As it is possible to observe by the top 10 words of each topic, the topics are coherent and specific.

In this, best result, the Topic 0 addresses the lack of knowledge and the challenges associated with transmitting information. Topic 1 focuses on public participation and involvement in policy-making. Topic 2 is about the scientific community. Topic 3 highlights communication barriers, Topic 4 reveals suggestions such as building networks to mediate researchers and politicians. Topic 5 is related to what constraints might lead to a bad collaboration between science and politics. Topic 6 discusses economic priorities, Topic 7 explores the intersection between science and policy. Topic 8 centers on media and communication channels, and finally, Topic 9 emphasizes language accessibility with the goal of making scientific communication more understandable (Table 4.6).

When examining these topics, it becomes evident that, although some thematic overlapping exists, for example, between Topic 2 and 7, where it is unclear whether the focus is on problems or solutions, the majority of topics are clear and coherent. Furthermore, since the data are retrieved from workshops specifically focused on the relationship between science and public policy, it is expected that similar themes would naturally recur across multiple topics.

TABLE 4.5. Singularity Score using BERTopic

Model	Preprocessing	Number of Topics	Singularity Score
AlBERTina	No Preprocessing	5	0.560
		10	0.556
		15	0.523
		20	0.514
	Simple Preprocessing	5	0.560
		10	0.544
		15	0.515
		20	0.532
	Normal Preprocessing	5	0.807
		10	0.852
		15	0.549
		20	0.536
	Portuguese Normal Preprocessing	5	0.807
		10	**0.936**
		15	0.563
		20	0.540
	Total Preprocessing	5	0.807
		10	0.484
		15	0.456
		20	0.470
	Portuguese Total Preprocessing	5	0.512
		10	0.512
		15	0.456
		20	0.444
Default	No Preprocessing	5	0.616
		10	0.560
		15	0.565
		20	0.572
	Simple Preprocessing	5	0.616
		10	0.616
		15	0.552
		20	0.828
	Normal Preprocessing	5	0.863
		10	0.584
		15	0.581
		20	0.842
	Portuguese Normal Preprocessing	5	0.552
		10	0.942
		15	0.871
		20	0.602
	Total Preprocessing	5	0.512
		10	0.504
		15	0.507
		20	0.480
	Portuguese Total Preprocessing	5	0.528
		10	0.536
		15	0.493
		20	0.490

TABLE 4.6. Top 10 topic words of the experiment with higher SS (0.936) - "Portuguese Normal" dataset, BERTopic model with Multilingual sentence-transformer and 10 topics

Topic 0	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**
falta	eventos	ciência	comunicação	stakeholders
conhecimento	decisores	científica	mensagem	atores
prazo	policy	investigação	canais	dados
resultados	pública	científico	comunicar	criar
informação	públicas	cientistas	dificuldade	redes
longo	políticas	investigadores	discurso	mapear
interesses	briefs	comunicação	falta	investigação
criar	público	sociedade	contacto	academia
complexidade	participar	conhecimento	forma	integrar
problemas	participação	falta	mensagens	brokers

Topic 5	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**
políticos	económicos	política	media	linguagem
agenda	prioridades	ciência	sociais	linguagens
política	económicas	políticos	redes	acessível
políticas	dinheiro	investigadores	social	diferentes
político	financiamento	políticas	meios	técnica
decisores	social	cientistas	mediática	pouco
interesses	objetivos	político	marketing	mesma
agendas	interesses	de	comunicação	simples
interesse	benefícios	decisores	jornalistas	comum
falta	económica	científicos	imprensa	públicos

On the other hand, there were two experiments with the lowest SS, both used the same model, LDA and number of topics generated, 5, with 0.4 of SS. The topics retrieved from the experiment with the no preprocessing dataset is shown in Table 4.7, and the one with the simple preprocessing dataset in Table 4.8. Analyzing Table 4.7 it is possible to understand that the topics are filled with punctuation and stop-words which does not convey any meaningful information. Topics are filled with noise due to the nature of dataset where no kind of preprocessing was applied. It is understandable that these are a really bad group of topics and it is impossible to take out any relevant information about the data.

Observing Table 4.8, and comparing with the previous experiment (Table 4.7), it is possible to conclude that the topics are better and with less noise. However, it is still hard to interpret the themes of each topic, since they are too noisy and their top words are mainly stop-words. Topic 0 could vaguely refer to communication/ language. Topic 1 has only two meaningful words: policies and communication, and they are not related enough to understand what the topic refers to. Topic 2, is slightly more interpretable, where the theme is the science-policy relation and its decision-making. Topic 3, seems to repeat the same theme of science-policy relations, yet those are the only two meaningful

TABLE 4.7. Top 10 topic words of the experiment with lowest SS (0.4) - "No Preprocessing" dataset, LDA model and 5 topics

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
de	e	a	de	de
;	Linguagem	de	(	,
Comunicação	com	e	)	comunicação
a	a	da	“	e
”	os	(	”	;
em	na	)	a	a
)	científica	ciência	para	)
(	stakeholders	para	com	(
“	não	política	e	os
com	sociedade	o	os	Falta

TABLE 4.8. Top 10 topic words of the experiment with lowest SS (0.4) - "Simple" dataset, LDA model and 5 topics

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
de	de	a	a	de
linguagem	a	criar	de	comunicação
a	e	os	ciência	e
e	com	de	o	falta
com	políticas	decisores	e	da
não	comunicação	para	para	políticos
do	o	com	da	ciência
redes	os	dos	do	dos
pouco	não	investigadores	política	na
criar	para	políticos	é	em

tokens. Finally, Topic 4, the cleanest one, is about the lack of communication between science and policymakers.

4.8. Topic Modeling with Large Language Model (LLM)

Beyond the traditional topic modeling techniques tested, such as LDA and BERTopic, less traditional but already really relevant, it was considered worthwhile to investigate how a Large Language Model LLM performs on typical NLP tasks, including topic modeling.

To do this, Google's Gemini API was used to group the textual suggestions into topics, give a possible title to that topic and show the top 10 significant words present in the documents that belong to that topic. The following text block includes the prompt used to retrieve the topics.

```
Below is a description of a text classification task, and some sug-
gestions in Portuguese. Your task is to group the suggestions into dif-
ferent topics based on their meaning.
```

Instructions:

- Each suggestion should appear in only one topic.
- Make as many groups as needed, based on the natural topics in the suggestions.
- For each topic, provide:
  - A possible title in English that summarizes the topic.
  - The top 10 significant words in Portuguese, drawn from the suggestions in that topic (words should reflect their importance or frequency within the topic, similar to TF-IDF weighting, but do not limit yourself strictly to a formula).

You'll be provided with the suggestions to group, each delimited by quotes: 'suggestion1', 'suggestion2', ...

Texts: {texts}

Two different models were used with the same prompt: Gemini 2.5 Pro [2] and Gemini 2.5 Flash [3]. The models returned a different number of topics, eight and seven topics, respectively. The names given to each one of the topics and the respective most important words are shown in Table 4.9 for the Gemini 2.5 Pro model and in Table 4.10 for the Gemini 2.5 Flash model.

By analyzing and comparing the tables, it is possible to identify many similarities between both results, the topics 1.1, 1.3 and 1.6 (Table 4.9) are identical to topics 2.1, 2.4 and 2.6 (Table 4.10), respectively.

When looking at the topics generated by both models, it can be observed that the topics are coherent, and the words are thematically related both to the overall subject and to each other within their topic. However, the last topic produced by the Gemini 2.5 Pro model appears weaker, as only three or four words meaningfully contribute to understanding the theme, while the remaining can be considered noisy due to their recurrence in multiple other topics. Therefore, Gemini 2.5 Flash appears to have performed better in this task. This conclusion is based only on qualitative observation of the topics and their top words, without the support of any quantitative metrics.

The increasing use of LLM to perform this type of task further reinforces the need of a metric that is able to evaluate quantitatively topics based on top words, with the ultimate goal of comparing models and selecting the most appropriate one for the task at hand. Although it was anticipated that the SS would not yield optimal results (given the characteristics of the words and the model, which tends not to produce multiple words from the same family due to their similar meanings), the SS was still tested on the topics generated by the Gemini models. Table 4.11 presents the SS results for both models. As previously noted, the results were not ideal, although Gemini 2.5 Flash retrieved a slightly better result, consistent with the qualitative observation.

[2] https://ai.google.dev/gemini-api/docs/models?hl=en#gemini-2.5-pro
[3] https://ai.google.dev/gemini-api/docs/models?hl=en#gemini-2.5-flash

TABLE 4.9. Top 10 topic words and titles of the model "Gemini 2.5 Pro"

Topic 1.1 Communication and Language Barriers	**Topic 1.2 Complexity and Nature of Scientific Knowledge**	**Topic 1.3 Disconnect Between Science and Policy/Society**	**Topic 1.4 Political, Economic, and Ideological Barriers**
comunicação	complexidade	ciência	políticos
linguagem	informação	sociedade	interesses
mensagem	conhecimento	realidade	agenda
cientistas	científico	academia	económicos
clara	resultados	falta	decisores
técnica	específica	integração	política
dificuldade	difícil	problemas	vontade
acessível	conteúdo	políticas	falta
discurso	teóricos	desconexão	poder
falar	dados	investigadores	conflito

Topic 1.5 Divergent Timelines and Priorities	**Topic 1.6 Solutions: Strategic Communication and Dissemination**	**Topic 1.7 Solutions: Building Bridges and Fostering Collaboration**	**Topic 1.8 Solutions: Capacity Building and Institutional Reform**
tempo	comunicação	criar	formação
prazo	mensagem	diálogo	capacitação
ciclo	policy	espaços	criar
político	briefs	parcerias	políticos
curto	linguagem	encontros	cientistas
longo	divulgação	stakeholders	investigadores
horizonte	estratégia	fóruns	políticas
prioridades	adaptar	redes	mecanismos
agenda	simplificar	participação	brokers
resultados	apresentar	eventos	avaliação

4.9. Discussion

The goal in Chapter 4 was to achieve a quantitative evaluation of the experiments made and, consequently, try to find the combination of preprocessing steps, number of topics generated, and model able to retrieve the most interpretable set of topics as possible.

After all the tests made to evaluate the quality of the topics yielded by each experiment, it becomes clear that this task is not as simple as it could seem since the more classical evaluation measures used were not clear or, even, informative at all. However, it is still possible to gain some knowledge from the results obtained.

Although the classical evaluation measures usually used for topic modeling show that LDA gains an advantage over BERTopic, in reality, both the human appreciation described in Section 3.4 and the SS proposed and calculated in Section 4.7.2 portray experiments using BERTopic are generally more interpretable and coherent.

BERTopic presents this advantage due to the characteristics of our data, composed with tiny texts and a limited number of documents, its own architecture better suited to capture contextual and semantic relationships within small datasets and the existence of

TABLE 4.10. Top 10 topic words and titles of the model "Gemini 2.5 Flash"

Topic 2.1 Communication Barriers and Clarity	**Topic 2.2 Relevance and Applicability Gaps**	**Topic 2.3 Political and Economic Drivers**	**Topic 2.4 Lack of Interaction and Knowledge Gaps**
linguagem	ciência	políticos	falta
comunicação	realidade	interesses	conhecimento
complexidade	objetivos	política	decisores
dificuldade	relevância	agenda	ciência
científica	problemas	económicos	políticos
técnica	aplicação	conflito	interação
clareza	necessidades	vontade	litercia
informação	investigação	prioridades	desconhecimento
simplificar	desconexão	curto	sensibilidade
mensagem	contexto	lobbies	confiança

Topic 2.5 Resource and Time Constraints	**Topic 2.6 Engagement Strategies & Communication Tools**	**Topic 2.7 Institutional Mechanisms & Collaboration**
financiamento	comunicação	criar
tempo	estratégia	decisores
falta	policy	cientistas
recursos	briefs	criação
custo	media	stakeholders
dinheiro	eventos	políticos
verbas	mensagem	espaços
orçamento	adaptar	integração
burocracia	formação	articulação
económicas	divulgação	mecanismos

TABLE 4.11. SS results for topics generated by "Gemini 2.5 Pro" and "Gemini 2.5 Flash"

Model	Singularity Score
Gemini 2.5 Pro	0.43
Gemini 2.5 Flash	0.48

Topic -1, which groups documents that do not clearly belong to any specific topic, as it helps prevent noise from affecting the coherence of the remaining topics. This means that even with the least preprocessed datasets, BERTopic preforms significantly better, when comparing SS results (Table 4.5 and Table 4.4).

Additionally, it is possible to conclude that the ideal number of topics for our data set lays between 5 and 10 since those are the experiments showing better results. Furthermore, it is not clear what the best preprocessing for the data is, albeit it seems to be between normal and normal translated since those experiments are the ones that are free of stopwords and use no lemmatization, which is valuable to BERTopic.

Finally, it is important to note that the LLM tested not only produced coherent and well-separated topics, but also generated meaningful labels for them, demonstrating potential for automating part of the topic modeling process, such as the creation of topic labels.

CHAPTER 5

Conclusions

This study focused on the text mining task, topic modeling, the data were tiny texts retrieved by PLANAPP's workshops, an initiative designed to bridge the gap between science and public policy. The main goal of this dissertation was to extract topics from these suggestions about public policies in Portuguese, as stated in **RQ1**. The literature review supported the decisions made throughout this work, particularly regarding the selection of topic modeling techniques. Although no studies were found with the exact same objective, related work dealing with non-English texts, short text data, or suggestion-based corpora provided valuable insights and reinforced the feasibility of this research. The objective was achieved, relevant and coherent topics were extracted, despite the inherent challenges associated with tiny textual data. Thus, it is possible to conclude that topic modeling techniques can be effectively applied to very short texts to extract meaningful insights, serving as a proof of concept for their applicability in this context.

In this study, we used two topic modeling techniques, LDA and BERTopic. With BERTopic, two different sentence-transformers, Multilingual and Portuguese pretrained, AlBERTina, were tested. Six datasets, each subjected to different preprocessing steps, were also utilized. In response to **RQ2**, which aimed to identify the most suitable modeling technique for our data, the results showed that, although the classical evaluation metrics tended to favor LDA, the topics generated by BERTopic were overall more coherent. However, the optimal sentence-transformer is not clear since its performance varied depending on the dataset.

The work in this research is characterized by an extensive number of experiments and, most importantly, by the difficulty of objectively evaluating the resulting topics. Throughout the study, several quantitative measures were used to evaluate the quality of the topics created, but none yielded results that we considered sufficiently robust. This limitation motivated the creation of the Singularity Score (SS), Section 4.7, an automated approach designed to approximate human evaluation and improve the assessment of topic coherence in tiny text scenarios.

5.1. Limitations and Future Work

Despite the successful extraction of coherent topics and the development of a new evaluation metric, this research presents several limitations. The main limitation concerns the data characteristics. Since the data were written in post-it notes, they are extremely short and limited in number, which restricts the representativeness of the corpus. In addition, the data being mainly in Portuguese but containing some records and technical words in English also presented a limitation of this work.

Another major limitation, closely related to the nature of the data, was the difficulty to evaluate the generated topics, as most metrics are not designed for tiny text. Therefore, SS was proposed, but it has limitations of its own. SS does not have an upper limit, meaning its values must be compared across experiments with the same data to retrieve valuable information. Furthermore, SS exclusively focus on words belonging to the same morphological family (words with the same stem). As a result, coherent and specific topics composed of synonyms or semantically related words may be incorrectly penalized.

Future work will therefore focus primarily on enhancing the SS. One direction involves integrating word embeddings or lexical databases to capture semantic similarity between words, allowing the metric to recognize coherent topics beyond morphological relations. This improvement could also mitigate another limitation of SS since it does not work for multilingual texts. Because stemmers are language-specific and word roots often differ across languages, the use of embeddings, and possibly multilingual embeddings, would make the metric more flexible and applicable to multilingual corpora. Lastly, future work should also involve further testing of SS in other datasets and languages and its validation with the help of annotators.

In summary, future research should focus on improving SS to make it a more robust and semantically informed evaluation metric. Such improvements would represent a valuable contribution to topic modeling, particularly when applied to tiny texts.

5.2. Scientific Contributions

This dissertation provides several contributions to topic modeling and, particularly, to tiny and Portuguese textual data. It presents an experimental study where two topic modeling techniques, LDA and BERTopic, and analyzes the strengths and weaknesses of both classical and transformer-based models, especially when applied to tiny texts. It also explores the use of an LLM to perform traditional text mining tasks, in this case, topic modeling.

Furthermore, Singularity Score (SS) is also introduced, a metric designed to mimic the behavior of annotators, enable the evaluation of the resulting topics regardless of record length. This is an important step toward automating and quantifying topic quality assessment.

In addition to these methodological and experimental contributions, a scientific poster and an extended abstract were also produced and presented for the *RECPAD 2025* (the annual Portuguese Conference on Pattern Recognition) [1]. The referred works are in Appendix B.

Finally, this study highlights the relevance and practical applications of topic modeling and it contributes to the broader goal of bridging the gap between science and public policy. Insights derived from the generated topics can help both parties identify weaknesses and pursue greater collaboration toward a more integrated and effective exchange of knowledge.

[1] `https://sites.google.com/view/recpad2025/home?authuser=0`

[This page is intentionally left blank.]

References

[1] D. Pati and L. N. Lorusso, "How to Write a Systematic Review of the Literature," en, *HERD: Health Environments Research & Design Journal*, vol. 11, no. 1, pp. 15–30, Jan. 2018, ISSN: 1937-5867, 2167-5112. DOI: 10.1177/1937586717747384. [Online]. Available: https://journals.sagepub.com/doi/10.1177/1937586717747384.

[2] A. Carrera-Rivera, W. Ochoa, F. Larrinaga, and G. Lasa, "How-to conduct a systematic literature review: A quick guide for computer science research," en, *MethodsX*, vol. 9, p. 101 895, 2022, ISSN: 22150161. DOI: 10.1016/j.mex.2022.101895. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2215016122002746.

[3] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd. 2025, Online manuscript released January 12, 2025. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/.

[4] L. Ma, S. Pahlevan Sharif, A. Ray, and K. W. Khong, "Investigating the relationships between MOOC consumers' perceived quality, emotional experiences, and intention to recommend: An NLP-based approach," *ONLINE INFORMATION REVIEW*, vol. 47, no. 3, pp. 582–603, May 2023. DOI: 10.1108/OIR-09-2021-0482.

[5] L. Hong, C. Fu, J. Wu, and V. Frias-Martinez, "Information Needs and Communication Gaps between Citizens and Local Governments Online during Natural Disasters," *INFORMATION SYSTEMS FRONTIERS*, vol. 20, no. 5, pp. 1027–1039, Oct. 2018. DOI: 10.1007/s10796-018-9832-0.

[6] H. Dang and J. Li, "Supply-demand relationship and spatial flow of urban cultural ecosystem services: The case of Shenzhen, China," *JOURNAL OF CLEANER PRODUCTION*, vol. 423, Oct. 2023. DOI: 10.1016/j.jclepro.2023.138765.

[7] V. D. H. De Carvalho and A. P. C. S. Costa, "Towards corpora creation from social web in Brazilian Portuguese to support public security analyses and decisions," en, *Library Hi Tech*, vol. 42, no. 4, pp. 1080–1115, Jul. 2024, ISSN: 0737-8831. DOI: 10.1108/LHT-08-2022-0401. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/LHT-08-2022-0401/full/html.

[8] J. Ningpeng, H. Tian, W. Haibo, X. Ruzhi, and M. Shiyu, "A Study on Structured Text Parsing for Policies Based on BERTopic," en, in *2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, Chongqing, China: IEEE, May 2024, pp. 16–22, ISBN:

979-8-3503-1653-7. DOI: 10.1109/IMCEC59810.2024.10575240. [Online]. Available: https://ieeexplore.ieee.org/document/10575240/.

[9] A. Lesnikowski, E. Belfer, E. Rodman, J. Smith, R. Biesbroek, J. D. Wilkerson, J. D. Ford, and L. Berrang-Ford, "Frontiers in data analytics for adaptation research: Topic modeling," *WILEY INTERDISCIPLINARY REVIEWS-CLIMATE CHANGE*, vol. 10, no. 3, Jun. 2019. DOI: 10.1002/wcc.576.

[10] K. Sprenkamp, M. Dolata, G. Schwabe, and L. Zavolokina, "Data-driven intelligence in crisis: The case of Ukrainian refugee management," English, *Government Information Quarterly*, vol. 42, no. 1, 2024, Publisher: Elsevier Ltd, ISSN: 0740624X (ISSN). DOI: 10.1016/j.giq.2024.101978. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85211342816&doi=10.1016%2fj.giq.2024.101978&partnerID=40&md5=1add767c5c9c4b352032ad30c4c436c7.

[11] K. Sprenkamp, L. Zavolokina, M. Angst, and M. Dolata, "Data-Driven Governance in Crises: Topic Modelling for the Identification of Refugee Needs," English, Association for Computing Machinery, 2023, pp. 1–11, ISBN: 979-840070837-4 (ISBN). DOI: 10.1145/3598469.3598470. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85167868516&doi=10.1145%2f3598469.3598470&partnerID=40&md5=ea06deb543b54c21516a58925f317126.

[12] R. Popping, "Analyzing open-ended questions by means of text analysis procedures," *Bulletin de méthodologie sociologique: BMS*, vol. 128, pp. 23–39, Oct. 2015. DOI: 10.1177/0759106315597389.

[13] G. Kalton and H. Schuman, "The Effect of the Question on Survey Responses: A Review," en, *Journal of the Royal Statistical Society. Series A (General)*, vol. 145, no. 1, p. 42, 1982, ISSN: 00359238. DOI: 10.2307/2981421. [Online]. Available: https://www.jstor.org/stable/2981421?origin=crossref.

[14] R. M. Silva, R. L. Santos, T. A. Almeida, and T. A. Pardo, "Towards automatically filtering fake news in portuguese," *Expert Systems with Applications*, vol. 146, p. 113 199, 2020, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2020.113199. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420300257.

[15] L. Hagen, "Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?" en, *Information Processing & Management*, vol. 54, no. 6, pp. 1292–1307, Nov. 2018, ISSN: 03064573. DOI: 10.1016/j.ipm.2018.05.006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306457317307240.

[16] R. Kowalski, M. Esteve, and S. Jankin Mikhaylov, "Improving public services by mining citizen feedback: An application of natural language processing," English, *Public Administration*, vol. 98, no. 4, pp. 1011–1026, 2020, Publisher: Wiley-Blackwell, ISSN: 00333298 (ISSN). DOI: 10.1111/padm.12656. [Online]. Available: https:

//www.scopus.com/inward/record.uri?eid=2-s2.0-85081980196&doi=10.1111%2fpadm.12656&partnerID=40&md5=b2f8f299dd36078222f8bb05df8017e4.

[17] L. Hagen, O. Uzuner, C. Kotfila, T. M. Harrison, and D. Lamanna, "Understanding Citizens' Direct Policy Suggestions to the Federal Government: A Natural Language Processing and Topic Modeling Approach," en, in *2015 48th Hawaii International Conference on System Sciences*, HI, USA: IEEE, Jan. 2015, pp. 2134–2143, ISBN: 978-1-4799-7367-5. DOI: 10.1109/HICSS.2015.257. [Online]. Available: http://ieeexplore.ieee.org/document/7070069/.

[18] A. Kousis and C. Tjortjis, "Investigating the Key Aspects of a Smart City through Topic Modeling and Thematic Analysis," *FUTURE INTERNET*, vol. 16, no. 1, Jan. 2024. DOI: 10.3390/fi16010003.

[19] J. Mazarura, "Topic modelling for short text," 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:62483684.

[20] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW '13, Rio de Janeiro, Brazil: Association for Computing Machinery, 2013, pp. 1445–1456, ISBN: 9781450320351. DOI: 10.1145/2488388.2488514. [Online]. Available: https://doi.org/10.1145/2488388.2488514.

[21] L. Hong, W. Yang, P. Resnik, and V. Frias-Martinez, "Uncovering topic dynamics of social media and news: The case of ferguson," in *Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part I*, Bellevue, USA: Springer-Verlag, 2016, pp. 240–256, ISBN: 978-3-319-47879-1. DOI: 10.1007/978-3-319-47880-7_15. [Online]. Available: https://doi.org/10.1007/978-3-319-47880-7_15.

[22] V. R. Sangaraju, B. K. Bolla, D. Nayak, and J. Kh, "Topic modelling on consumer financial protection bureau data: An approach using bert based embeddings," *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pp. 1–6, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248811133.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. null, pp. 993–1022, Mar. 2003, ISSN: 1532-4435.

[24] K. Porter, "Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling," *DIGITAL INVESTIGATION*, vol. 26, S87–S97, Jul. 2018. DOI: 10.1016/j.diin.2018.04.023.

[25] W. Zhao, J. J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, and W. Zou, "A heuristic approach to determine an appropriate number of topics in topic modeling," *BMC Bioinformatics*, vol. 16, no. 13, S8, Dec. 2015, ISSN: 1471-2105. DOI: 10.1186/1471-2105-16-S13-S8. [Online]. Available: https://doi.org/10.1186/1471-2105-16-S13-S8.

[26] Y. Zhou, L. Yan, and X. Liu, "A quantitative study of disruptive technology policy texts: An example of China's artificial intelligence policy," *JOURNAL OF DATA AND INFORMATION SCIENCE*, vol. 9, no. 3, pp. 155–180, Jun. 2024. DOI: 10.2478/jdis-2024-0016.

[27] T. Papadopoulos and Y. Charalabidis, "What do governments plan in the field of artificial intelligence?: Analysing national AI strategies using NLP," English, Association for Computing Machinery, 2020, pp. 100–111, ISBN: 978-145037674-7 (ISBN). DOI: 10.1145/3428502.3428514. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85095968003&doi=10.1145%2f3428502.3428514&partnerID=40&md5=f1162f07e869d35b8808009d45fd8120.

[28] Y. Chen and Chenyongjun Ding, "Multidimensional evolutionary analysis of China's BIM technology policy based on quantitative mapping," *ARCHITECTURAL ENGINEERING AND DESIGN MANAGEMENT*, vol. 20, no. 3, pp. 578–595, May 2024. DOI: 10.1080/17452007.2023.2291585.

[29] M. E. Roberts, B. M. Stewart, and D. Tingley, "Stm: An r package for structural topic models," *Journal of Statistical Software*, vol. 91, no. 2, pp. 1–40, 2019. DOI: 10.18637/jss.v091.i02. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v091i02.

[30] M. Ovadek, A. Dyevre, and K. Wigard, "Analysing EU Treaty-Making and Litigation With Network Analysis and Natural Language Processing," *FRONTIERS IN PHYSICS*, vol. 9, May 2021. DOI: 10.3389/fphy.2021.657607.

[31] S. Eckhard, R. Patz, M. Schönfeld, and H. van Meegdenburg, "International bureaucrats in the UN Security Council debates: A speaker-topic network analysis," English, *Journal of European Public Policy*, vol. 30, no. 2, pp. 214–233, 2023, Publisher: Routledge, ISSN: 13501763 (ISSN). DOI: 10.1080/13501763.2021.1998194. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119296832&doi=10.1080%2f13501763.2021.1998194&partnerID=40&md5=2a752592eea74e0290105a3d83d866c7.

[32] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in sociology*, vol. 7, p. 886 498, 2022.

[33] Z. Wang, J. Chen, J. Chen, and H. Chen, "Identifying interdisciplinary topics and their evolution based on bertopic," *Scientometrics*, vol. 129, Jul. 2023. DOI: 10.1007/s11192-023-04776-5.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423/.

[35] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, pp. 267–276, 1953. [Online]. Available: https://api.semanticscholar.org/CorpusID:120467216.

[36] J. MacQueen, "Multivariate observations," in *Proceedings ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability*, vol. 1, 1967, pp. 281–297.

[37] M. Syakur, B. Khusnul Khotimah, and E. Rohman, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," *IOP Conference Series: Materials Science and Engineering*, vol. 336, p. 012 017, Apr. 2018. DOI: 10.1088/1757-899X/336/1/012017.

[38] P. Rousseeuw, "Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987. DOI: 10.1016/0377-0427(87)90125-7.

[39] D. SAPUTRA, D. Saputra, and L. Oswari, "Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method," Jan. 2020. DOI: 10.2991/aisr.k.200424.051.

[40] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988, ISSN: 0306-4573. DOI: https://doi.org/10.1016/0306-4573(88)90021-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0306457388900210.

[41] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, pp. 43–52, Dec. 2010. DOI: 10.1007/s13042-010-0001-0.

[42] H. Trenquier, "Improving semantic quality of topic models for forensic investigation," *University of Amsterdam*, pp. 2017–2018, 2018.

[43] P. Rousseeuw, "Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987. DOI: 10.1016/0377-0427(87)90125-7.

[44] P. Xie, D. Yang, and E. Xing, "Incorporating Word Correlation Knowledge into Topic Modeling," en, in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado: Association for Computational Linguistics, 2015, pp. 725–734. DOI: 10.3115/v1/N15-1074. [Online]. Available: http://aclweb.org/anthology/N15-1074.

[This page is intentionally left blank.]

APPENDIX A

Complementary visualizations

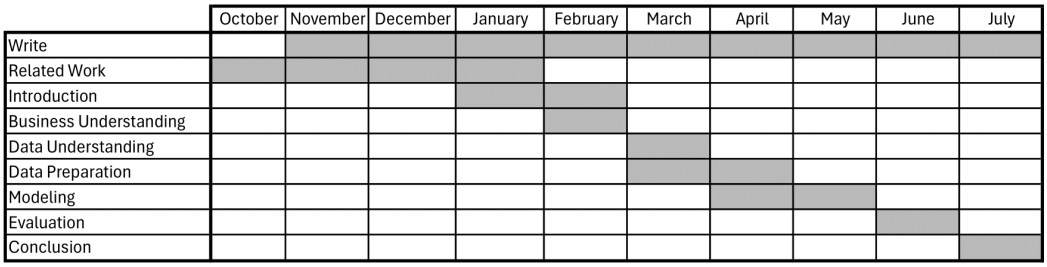


FIGURE A.1. Gantt Diagram of the thesis work

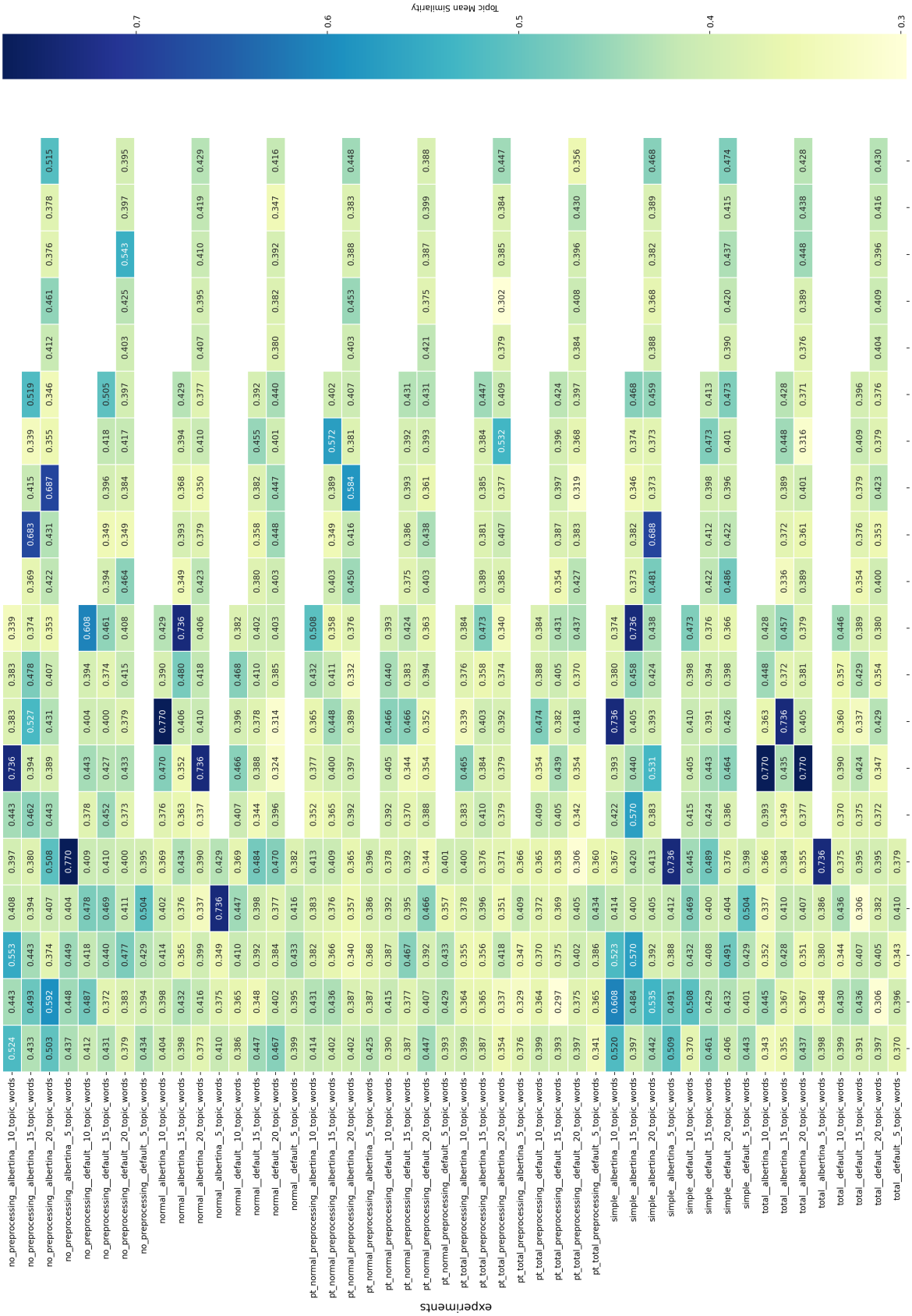


FIGURE A.2. Full heatmap of mean cosine-similarity between embeddings belonging to the same topic

APPENDIX B

Poster and extended abstract presented at RECPAD 2025

Singularity Score for Evaluating Topic Relevance in Tiny Text

Nicole Nunes
Nicole_Nunes@iscte-iul.pt
Ana Rita Peixoto
Rita_Peixoto@iscte-iul.pt
Ana Maria de Almeida
Ana.Almeida@iscte-iul.pt

Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal

Abstract

Topic modeling is a challenging task on its own, and when the text being analyzed is very short, the challenge increases significantly, as there are no reliable metrics to evaluate the quality of the generated topics. In this study, several experiments are conducted using different preprocessing methods and various topic modeling techniques to determine which approach produces the best results. The singularity score is also proposed as a metric to evaluate the quality of the topics created.

1 Introduction

Topic modeling is particularly relevant within the world of data mining and text mining to recognize patterns and extract themes from user-reported issues. The data used in this experiment were retrieved from "PLANAPP - Centro de Planeamento e Avaliação de Políticas Públicas" [1] (Center for Planning and Evaluation of Public Policies) workshops, whose objective is to support the development of evidence-informed policies and raise scientific awareness of this need. During the workshops, using post-it notes, participants wrote challenges and possible solutions to bridge the gap between science and policy, which PLANAPP then analyses and determines what are the main challenges or the main expected solutions. This entity needs an automated system able to determine the main topics derived from the workshops, i.e., it needs a topic modeling tool.

1.1 Topic Modeling: Methods and Challenges

When applying topic modeling, several factors must be considered. It is essential to account for the specificities of the data, as these may have significant implications for preprocessing and modeling. One of the issues here comes from having a two-language corpus (Portuguese and English), but two stand out: translating the documents written in the language with the least occurrences to the second, or employing a model specifically designed for multilingual settings such as BERTopic multilingual.

Survey responses can also be a challenge to topic modeling due to their informal style and short length, which means that they may not contain enough meaningful words [4]. The authors of [2, 3] chose to simply avoid the problem by removing documents with fewer words than a certain minimum. Others chose to apply models more appropriate to this text length, such as Biterm Topic Model [7], ST-LDA, Latent Dirichlet Allocation [1] and BERTopic [5]. Figure 1 shows the evolution of the topic modeling techniques applied in 18 documents that were analyzed. It is noticeable that the use of BERTopic has grown in recent years, although there is still a prevalence of classic methods such as LDA.

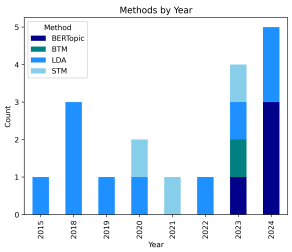


Figure 1: Topic Modeling Method used by Year

[1] https://www.planapp.gov.pt

2 Development

2.1 Methodology

After understanding the data, the next step was data preparation, then modeling, and finally evaluation of the results. These steps are iterative, that is, the modeling process does not exclude returning to the data preparation stage and the same goes for evaluation.

2.2 Data Understanding and Preparation

The data is composed of tiny phrases extracted from workshop post-its written in two languages, Portuguese and English, shows multiple abbreviations, is divided into six different Excel sheets and it is unlabeled. The data were collapsed into a single corpus, having a total of 1810 documents, and normalized afterwards by creating a dictionary which mapped abbreviations to their corresponding full words. To address the presence of technical words in English, their Portuguese translation was also incorporated into the dictionary.

Six data sets were created. The first involved only tokenization with no additional preprocessing steps involved, it had a 5.88 words per document on average. The second had simple preprocessing, where the punctuation and numbers were removed, the abbreviation dictionary was applied and all words were converted to lowercase, 5.89 words per document on average. The third used normal preprocessing which extended simple preprocessing by also removing stop words, 4.08 words per document on average. The fourth is a variant of this dataset and it was produced by translating the documents in English to Portuguese, having on average 4.12 words per document. The fifth and sixth employed total preprocessing - original and translated, adding lemmatization to the normal preprocessing, with a translated version of this dataset also being generated, having 4.08 and 4.13 average words per document, respectively.

2.3 Modeling

The first step implies deciding the number of topics that the model should create. This usually can be achieved by using the Elbow method [6]. An attempt was made to estimate the optimal number of topics using the elbow method, by calculating inertia and silhouette scores for 3 to 35 topics. Although these are classic methods and they usually work, as the average number of words in each document is less than 6, and there are only 1810 documents, these metrics usually perform worse in shorter text. As no secure conclusion could be drawn from the results, the decision was to test with 5, 10, 15, and 20 topics for each experiment.

In the modeling step itself, besides a Gemini request that works as a baseline, two methods were tested: the classic LDA and BERTopic with two different sentence transformers AlBERTina and Multilingual. Therefore, a total of 72 experiments were performed.

2.4 Evaluation

The objective is to determine which of the combinations performs best. Specifically, the aim is to ensure that the documents within the topic are sufficiently similar to belong together while, at the same time, are distinct enough from documents in other topics so that the topics are well separated and unambiguous.

Classic methods, such as Coherence and Perplexity are used to evaluate the quality of the topics created. Although coherence values above 0.8 are typically considered indicative of good results, none of the experiments reached values greater than 0.43. Like Coherence, Perplexity values were not ideal to conclude on which experiment provided the best topics. This outcome can be explained by the sparsity of the data, since such

experiments are usually performed with larger datasets and with longer documents than those used in this study.

Since the classical methods did not yield conclusive results, alternative approaches were explored, including: the application of the Tsallis and Rényi entropies, the calculation of the Silhouette Score based on embeddings of the representative words of each topic, and the computation of the mean cosine similarity of those same embeddings. However, it still was not possible to reach definitive conclusions.

It is important to note that these metric results do not necessarily imply the topics are poorly formed or the models are incapable of identifying them. Rather, they indicate that the metrics themselves are not well suited to the type of data used in this study. Consequently, to determine which experiment produces better topics, it is common practice to rely on human annotators to visually assess whether the results are coherent. Although human annotation is a common practice, it is interesting to explore a quantitative approach to reliably evaluate topics, even when these are derived from texts that are typically more challenging. With this need in mind, we have developed the Singularity Score described next.

Singularity Score

The premise of Singularity Score is to assess the quality of topics based on the top 10 most significant words in each, thus using the top 10 word stems (word roots) as a starting point. Assuming the model has N topics and for each topic i, Significant Word, $SW \in \{0,1\}$, Count of Unique Words, $UW \in [0,10]$ and Count of Non-Unique Words, $NUW \in [0,10]$

If at least one stem appears more than once in the same topic (SW) the score is increased. Then the within-topic consistency is evaluated: if the SW is unique to the topic (UW), this means that it is not repeated in another topic and the score increases. If, on the other hand, the SW is repeated in another topic (NUW), the score decreases. This calculation is called Topic Uniqueness (tu_i) and is calculated for every topic in the experience using the equation 1.

$$tu_i = w_{SW} SW_i + w_{UW} \frac{UW_i}{10} + w_{NUW} \left(1 - \frac{NUW_i}{10}\right), w_{SW} + w_{UW} + w_{NUW} = 1 \tag{1}$$

To get the TU of an experiment, the average of the values for each topic is used. However, this approach may be excessively punitive since, even when the majority of topics demonstrate high quality, the presence of a small number of weaker topics can substantially reduce the average. Therefore, a reward was placed on experiments that produced a significant proportion of well-performing topics (θ). ST is the fraction of Strong Topics given by equation 2. Then, equation 3, applied to ST, rewards models that have more than half of the topics with strong topics (high values of tu). The β value is the relative bonus ($\beta \in [0,1]$). Singularity Score is determined after rewarding TU, presented in equation 4.

$$ST = \frac{number\ of\ topics\ with\ tu \geq \theta}{N} \tag{2}$$

$$f(ST) = \begin{cases} 0 & if\ ST \leq 0.5 \\ \beta ST & if\ ST > 0.5 \end{cases} \tag{3}$$

$$SS = TU(1 + ST) \tag{4}$$

2.5 Results

Figure 2 shows the topics of the best scoring experience using Singularity Score, which was the translated text using normal preprocessing, AlBERTina sentence-transformer, and 5 topics. As can be observed, each topic has a specific theme: Topic 1 is about Language and Communication, Topic 2 is related to Politics, Topic 3 reveals Research and Projects, Topic 4 is about Goals and Deadlines, and Topic 5 Science, Politics, and Gaps in knowledge transfer.

3 Conclusions and Future Work

This study brings out the challenges issuing when addressing Topic Modeling for handling extremely short text, where applying traditional metrics prove to be insufficient. To address this limitation, the Singularity Score metric was introduced (Subsection 2.4). Our experimentation provided evidence that this is an effective measure for topic relevance.

Figure 2: Word Clouds of an Experiment with SS of 0.998

Future work will focus on validating Singularity Score with classical datasets to address whether its performance remains consistent. In addition, exploration will be conducted on the weights assigned to the metric.

This work was partially supported by Fundação para a Ciência e a Tecnologia, I.P. (FCT) [Project 2024.07395.IACDC][ISTAR Projects: UIDB/04466/2023 and UIDP/04466/2023]

References

[1] Lingzi Hong, Weiwei Yang, Philip Resnik, and Vanessa Frias-Martinez. Uncovering topic dynamics of social media and news: The case of ferguson. In *Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part I*, page 240–256, Berlin, Heidelberg, 2016. Springer-Verlag. ISBN 978-3-319-47879-1. doi: 10.1007/978-3-319-47880-7_15. URL https://doi.org/10.1007/978-3-319-47880-7_15.

[2] Lingzi Hong, Cheng Fu, Jiahui Wu, and Vanessa Frias-Martinez. Information Needs and Communication Gaps between Citizens and Local Governments Online during Natural Disasters. *INFORMATION SYSTEMS FRONTIERS*, 20(5):1027–1039, October 2018. doi: 10.1007/s10796-018-9832-0.

[3] Lan Ma, Saeed Pahlevan Sharif, Arghya Ray, and Kok Wei Khong. Investigating the relationships between MOOC consumers' perceived quality, emotional experiences, and intention to recommend: an NLP-based approach. *ONLINE INFORMATION REVIEW*, 47(3):582–603, May 2023. doi: 10.1108/OIR-09-2021-0482.

[4] Jocelyn Mazarura. Topic modelling for short text. 2015. URL https://api.semanticscholar.org/CorpusID:62483684.

[5] Vasudeva Raju Sangaraju, Bharath Kumar Bolla, Deepa Nayak, and Jyothsna Kh. Topic modelling on consumer financial protection bureau data: An approach using bert based embeddings. *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pages 1–6, 2022. URL https://api.semanticscholar.org/CorpusID:248811133.

[6] Robert L. Thorndike. Who belongs in the family? *Psychometrika*, 18:267–276, 1953. URL https://api.semanticscholar.org/CorpusID:120467216.

[7] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 1445–1456, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320351. doi: 10.1145/2488388.2488514. URL https://doi.org/10.1145/2488388.2488514.