FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Exploring Entity and Topic Extraction in Portuguese News

Marcos Rafael Peixoto Aires

Master's Degree in Informatics and Computing Engineering

Supervisor: Prof. Sérgio Nunes

July 7, 2025

Exploring Entity and Topic Extraction in Portuguese News

Marcos Rafael Peixoto Aires

Master's Degree in Informatics and Computing Engineering

Approved in oral examination by the committee:

President: Prof. João Correia Lopes
Referee: Prof. Nuno Escudeiro
Referee: Prof. Sérgio Nunes

July 7, 2025

Resumo

Nos últimos anos, as notícias online tornaram-se uma das principais fontes de informação para a população portuguesa, criando novas oportunidades e desafios para a compreensão das dinâmicas do jornalismo digital. Esta dissertação investiga como é que técnicas de Processamento de Linguagem Natural (PLN) podem ser aplicadas para extrair informações relevantes — nomeadamente entidades mencionadas e tópicos — a partir de conteúdos jornalísticos portugueses disponíveis online. O principal objetivo é desenvolver e avaliar metodologias que permitam identificar padrões nas dinâmicas de publicação noticiosa e na cobertura de conteúdos por diferentes meios de comunicação em Portugal.

Para tal, foi concebido um sistema completo que processa uma grande coleção de títulos e resumos de notícias recolhidos de várias fontes online. A cadeia de processamento desenvolvida inclui a caracterização da coleção, modelação e etiquetagem de tópicos, bem como Reconhecimento de Entidades Mencionadas (REM). Diversas ferramentas e modelos foram analisados e comparados quanto à sua eficácia nestas tarefas. No caso de modelação de tópicos, BERTopic, Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM), e Latent Dirichlet Allocation (LDA) foram avaliados, ao passo que no caso de etiquetagem de tópicos, a ferramenta Yake! foi testada. No que toca a REM, spaCy, Generalist Model for Named Entity Recognition using Bidirectional Transformer (GLiNER), e um modelo baseado em BERTimbau foram avaliadas. As ferramentas de REM e as de modelação de tópicos foram avaliadas com base em métricas quantitativas — pontuação F1 para ferramentas de REM, e Taxa de acertos para ferramentas de modelação de tópicos. No caso das ferramentas de modelação e etiquetagem de tópicos, estas foram também avaliadas através de inquéritos qualitativos com participantes humanos.

A informação extraída — composta por tópicos e entidades — foi integrada numa base de dados, e, através do uso de uma ferramenta de visualização de dados — Metabase — foram ainda desenvolvidas estratégias de visualização com o objetivo de fazer uma análise comparativa dos padrões de publicação e da diversidade de cobertura entre diferentes fontes em diferentes granularidades temporais.

Os resultados revelam diferenças na forma como os meios de comunicação portugueses abordam tópicos e entidades ao longo do tempo, e demonstram o valor de combinar técnicas automatizadas com avaliações centradas no utilizador para a análise do ecossistema noticioso português. Por um lado, os profissionais dos media podem beneficiar destes resultados para ajustar estratégias editoriais, identificar tendências emergentes e monitorizar as dinâmicas de publicação de outras fontes noticiosas. Por outro, os resultados apresentados abrem caminho para investigações futuras no campo dos estudos jornalísticos, ao fornecer uma base sólida para a exploração dos processos de produção noticiosa e da evolução temática no ecossistema digital português.

Palavras-chave: Processamento de Linguagem Natural, Reconhecimento de Entidades Nomeadas, Texto Jornalístico

Abstract

In recent years, online news has become one of the main sources of information for the Portuguese population, creating new opportunities and challenges for understanding the dynamics of digital journalism. This dissertation investigates how Natural Language Processing (NLP) techniques can be applied to extract relevant information — named entities and topics — from Portuguese journalistic content available online. The main objective is to develop and evaluate methodologies that enable the identification of patterns in news publication dynamics and content coverage across different media outlets in Portugal.

To this end, a complete system was designed to process a large collection of news headlines and summaries gathered from various online sources. The developed processing pipeline includes collection characterization, topic modeling and labeling, as well as Named Entity Recognition (NER). Several tools and models were analyzed and compared in terms of their effectiveness in these tasks. For topic modeling, BERTopic, Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM), and Latent Dirichlet Allocation (LDA) were evaluated, while for topic labeling, the Yake! tool was tested. Regarding NER, spaCy, the Generalist Model for Named Entity Recognition using Bidirectional Transformer (GLiNER), and a model based on BERTimbau were evaluated. The NER and topic modeling tools were evaluated using quantitative metrics — F1-score for NER tools, and accuracy rate for topic modeling tools. Additionally, topic modeling and labeling tools were evaluated through qualitative surveys involving human participants.

The extracted information — composed of topics and entities — was integrated into a database, and, using a data visualization tool — Metabase — visualization strategies were developed to enable a comparative analysis of publication patterns and coverage diversity across different sources and temporal granularities.

The results reveal differences in how Portuguese media outlets approach topics and entities over time and demonstrate the value of combining automated techniques with user-centered evaluations for analyzing the Portuguese news ecosystem. On the one hand, media professionals can benefit from these findings to adjust editorial strategies, identify emerging trends, and monitor the publication dynamics of other news sources. On the other hand, the results presented pave the way for future research within the field of journalism studies, by providing a solid foundation for exploring news production processes and thematic evolution in the Portuguese digital ecosystem.

Keywords: Natural Language Processing, Named Entity Recognition, News Text

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework to achieve a better and more sustainable future for all. It includes 17 goals to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice.

This dissertation, which focuses on Natural Language Processing (NLP) techniques for analyzing Portuguese digital news, is particularly aligned with SDGs related to education, innovation, and information transparency. It demonstrates how interdisciplinary approaches can help monitor public discourse, analyze media diversity, and enable data-driven decision-making.

The specific Sustainable Development Goals mentioned have the following names:

SDG 4 Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation

SDG 16 Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels

SDG	Target	Contribution	Performance Indicators and Metrics
4	4.7	Fosters media literacy and promotes learning through technology. Contributes to educational goals by supporting research and learning in NLP, journalism studies, and information systems.	• Availability of documented methodologies and models. • Number of educational applications (e.g., academic references, tools used in courses). • Dissemination of visual analytics and interpretability resources.
9	9.5	Contributes to enhancing scientific research and innovation through the evaluation of NLP methodologies for Portuguese digital journalism. Supports the creation of intelligent systems to process and analyze large-scale media content.	• Number and type of NLP models evaluated (BERTopic, GSDMM, LDA, spaCy, GLiNER, BERTimbau-based). • Model performance metrics (F1-score, hit rate, qualitative user evaluation). • Framework for topic/entity extraction and analysis.
16	16.10	Enables transparent access to media dynamics, promoting the monitoring of information diversity and editorial trends. Empowers both researchers and the public to assess how institutions (media) represent societal topics and entities.	• Diversity measures in topic and entity coverage per media outlet. • Visualization dashboards using Metabase. • Number of news sources processed and analyzed.

Acknowledgements

I could not stress more how grateful I am for everyone who helped me through this very long journey. I may admit this was the biggest challenge I have ever had to conquer, from the long nights with almost no sleep to the endless days when the only question on my mind was: "Will I be able to do this?". And after nine long months — defined by one of the most important words in my life: perseverance — I can finally say that I was able to do it.

First, I want to thank my supervisor, Sérgio Nunes, for all the help he gave me and for never giving up on me. From the weekly meetings to the constant support on Slack, he was always tireless when it came to assisting me. Without him, this would not be possible.

I would also like to thank my family for always believing in me and for being supportive, mainly my parents and my grandparents. Thank you for always investing in my education and for always instilling in me the importance of being resilient, even in the face of obstacles that seem insuperable. If I am who I am today, it is because of you.

I extend my heartfelt gratitude to all my friends, who have always stood by my side. Thank you for your kind words, valuable advice, support throughout the dissertation process, and the countless moments of uncontrollable laughter you never fail to bring. You can always put me in a good mood, even when the world seems to be falling apart. You all hold a special place in my heart.

Finally, I want to express my gratitude to my girlfriend. When I was as close to giving up as I could be, you always reminded me to keep my head up and to remember that, if I was capable of getting this far, I was also able to finish this. Thank you for always listening to my complaints, for the walks by the beach that always cleared my mind, and for always believing in me, even more than I believed in myself at times. Without your motivation, this would not be possible.

Without each of you, I am 100% sure I would not have had the strength to complete this dissertation. This work is as much yours as it is mine.

Marcos Aires

"Há tempo para rodar, correr no mundo.
Há tempo para voar e bater no fundo.
Mas o tempo está a andar e o tempo é curto.
Já é tempo para ser adulto."

Allen Pires Sanhá

Contents

Resumo i

Abstract ii

1 Introduction 1
- 1.1 Context and Motivation . . . 1
- 1.2 Problem . . . 2
- 1.3 Research Questions . . . 3
- 1.4 Document Structure . . . 3

2 Literature Review 4
- 2.1 Identification Phase . . . 4
  - 2.1.1 Refinement Phase . . . 6
  - 2.1.2 Final Search Strings . . . 6
  - 2.1.3 Search criteria . . . 7
- 2.2 Screening Phase . . . 7
  - 2.2.1 Inclusion and Exclusion Criteria . . . 7
  - 2.2.2 Removed duplicates . . . 8
  - 2.2.3 Screened records . . . 8
  - 2.2.4 Forward and Backward Snowballing . . . 8
- 2.3 Eligibility Phase . . . 8
  - 2.3.1 First Eligibility Phase . . . 8
  - 2.3.2 Second Eligibility Phase . . . 9
- 2.4 Summary . . . 10

3 State of the Art 11
- 3.1 Variations in Coverage . . . 11
  - 3.1.1 Topic Coverage . . . 11
  - 3.1.2 Entity Coverage . . . 17
- 3.2 News Publication Patterns . . . 18
  - 3.2.1 Publication and Update Patterns . . . 18
- 3.3 Media Diversity . . . 21
  - 3.3.1 General Overview . . . 21
  - 3.3.2 Frameworks for Media Diversity . . . 22
  - 3.3.3 Media Diversity Evaluation . . . 22
- 3.4 Summary . . . 27

4 Solution **29**
4.1 Database Creation . . . 29
4.2 Corpus Characterization . . . 30
4.3 Corpus Evaluation . . . 33
4.4 Topic and Entity Extraction . . . 34
4.4.1 Topic Modeling . . . 35
4.4.2 Topic Labeling . . . 37
4.4.3 Named Entity Recognition . . . 40
4.5 Summary . . . 45

5 Results and Discussion **46**
5.1 User survey and Entity Extraction Tools Evaluation Results . . . 46
5.2 Dataset Enrichment with Extracted Topics and Entities . . . 49
5.3 Metabase . . . 52
5.4 Publication Patterns . . . 53
5.5 Variations in Coverage . . . 54
5.5.1 Entity Coverage . . . 55
5.5.2 Entity Type Coverage . . . 57
5.5.3 Topic Coverage . . . 59
5.6 Summary . . . 63

6 Conclusions and Future Work **65**
6.1 Final Remarks . . . 65
6.2 Future Work . . . 67

References **69**

List of Figures

2.1 Literature Review steps flowchart 5

3.1 Visualization example of the "Keywords tool", showing the percentage of articles published about each of the topics in news sources [25]. 13
3.2 Visualization example of the "Keywords tool", showing the percentage of articles published about each of the topics in blog sources [25]. 14
3.3 Time-sets visualization showing the amount of attention given to each of the topics mentioned across three different news publication sources [46]. 14
3.4 Difference in percentage of political articles posted between election and routine periods [22]. 15
3.5 Topic coverage between different geographical levels [64]. 16
3.6 Coverage on cohesion policy over time in different geographical contexts [64]. . 16
3.7 Visualization example of the "NewsMap tool", showing the coverage the term "election" received worldwide [26]. 17
3.8 Example of visualization of the "Sources tool", showing the percentage of articles published by each of the sources [25]. 19
3.9 Temporal patterns of content change rate for the MSNBC news webiste [14]. . . 20
3.10 Prediction of the change patterns over three different time horizons (1, 3, and 6 hours) [14]. 20
3.11 Diversity over time for windows of 5 days in the month of September [25]. . . . 26
3.12 tweet-word matrix [37]. 27

4.1 Database schema represented through a UML diagram. 30
4.2 News per Portuguese source from January to October 2024. 33
4.3 Question to evaluate the topic modeling tools. 39
4.4 Question to evaluate the topic labeling strategies. 40

5.1 User survey results for one of the LDA clusters. 47
5.2 User survey results for one of the GSDMM clusters. 48
5.3 Articles published for each month of the year (%). 53
5.4 Articles published for each day of the month (%). 54
5.5 Articles published for each day of the week (%). 54
5.6 Articles published during a 24-hour cycle (%). 54
5.7 Types of entities mentioned in articles published in August 12th. 55
5.8 Percentage of articles mentioning "Cristiano Ronaldo" (represented by the purple line) and "Donald Trump" (represented by the green line) by month in 2024. . . . 56
5.9 Average number of entities mentioned per article, per source. 56
5.10 Most mentioned entities in the whole dataset. 57

5.11 Most mentioned entities per source, considering the five sources with the highest average number of entities mentioned per article. 57
5.12 Percentage of articles mentioning "PER" entity types (represented by the red line) and "LOC" entity types (represented by the yellow line) by month in 2024. 57
5.13 Entity types ordered by number of mentions in the whole dataset. 58
5.14 Entity types ordered by number of mentions per source. 58
5.15 Number of topics covered per source by BERTopic. 59
5.16 Number of topics covered per source by GSDMM. 59
5.17 Average number of topics covered per article, per source considering BERTopic modeling. 60
5.18 Average number of topics covered per article, per source, considering GSDMM modeling. 60
5.19 Percentage of articles mentioning "morreu, gregory, alf" (represented by the blue line) and "irs, jovem" (represented by the orange line) by month in 2024 in the case of BERTopic. 61
5.20 Percentage of articles mentioning "morreu, gregory, alf" (represented by the blue line) and "irs, jovem" (represented by the orange line) by month in 2024 in the case of GSDMM. 61
5.21 Most covered topics in the whole dataset in the case of BERTopic. 62
5.22 Most covered topics per source in the case of BERTopic, considering the three sources with the highest average number of topics mentioned per article and two other sources. 62
5.23 Most covered topics in the whole dataset in the case of GSDMM. 63
5.24 Most covered topics per source in the case of GSDMM, considering the three sources with the highest average number of topics mentioned per article and two other sources. 63

List of Tables

2.1 Selected records after the second eligibility phase 9

3.1 Topic saliency and Shannon's H result for each category [4] 24

4.1 Sources per type and news per source type, in absolute numbers and percentages 33
4.2 Hit rate for each of the topic modeling tools. 38
4.3 Entities correspondence between spaCy, GLiNER, and the BERTimbau-based model. 42
4.4 Rules for identification and labeling of the different entity types [35, 84]. 44
4.5 Evaluation of one news document with number of "TP", "FP", and "FN". 44

5.1 F1-score for each of the entity extraction tools (first evaluation). 46
5.2 F1-score for each of the entity extraction tools (second evaluation). 47
5.3 Scores for each topic modeling tool obtained by the calculation of the weighted mean of the evaluations given by the participants. 48
5.4 Scores for each topic labeling strategy obtained by the calculation of the weighted mean of the evaluations given by the participants. 49

Listings

4.1 Example entry from the `articles` table. 31
4.2 Example entry from the `cats` table. 31
4.3 Example entry from the `articles_cats` table. 31
4.4 Example entry from the `sources` table. 31
4.5 Example entry from the `feeds` table. 32
4.6 Example of a news document from the JSON file with entities extracted by all the tools. 42
5.1 Example of a news document from the JSON file with entities extracted by the final tools. 50
5.2 Example entry from the `articles_everything` table. 51

Abreviaturas e Símbolos

NLP	Natural Language Processing
DL	Deep Learning
ML	Machine Learning
LDA	Latent Dirichlet Allocation
STM	Structural Topic Model
dDTM	discrete Dynamic Topic Model
NER	Named Entity Recognition
GLiNER	Generalist Model for Named Entity Recognition using Bidirectional Transformer
GSDMM	Gibbs Sampling Dirichlet Multinomial Mixture
DTM	Dynamic Topic Model

Chapter 1

Introduction

Online news has become part of people's everyday lives worldwide, emerging as one of the primary means of disseminating information, and Portugal is no exception to the rule. According to a study made by OberCom, in 2023, an estimated 73.6% of Portuguese citizens used the Internet to access news, while 37.8% relied on it as their primary source of news [67]. Social media also contributes to this high percentage, as they are now channels for online news dissemination. According to the same study, as of 2023, 10.9% of Portuguese people paid for news in digital format, a fact that shows that a considerable amount of people find it worthwhile to pay for this type of service. In fact, according to OberCom, 40% of people reported that they pay for online news because they feel they benefit from content they cannot get from other sources, and 36.5% said that they feel they obtain better quality by paying than by using any free source. These findings underscore the central role of online news in the daily lives of Portuguese citizens, both as a widely accessed information source and as a service many are willing to pay for in exchange for exclusive or higher-quality content.

1.1 Context and Motivation

The analysis of the online news ecosystem is facilitated by integrating insights from multiple areas of informatics. Indeed, various informatics domains have been employed in prior research to explore the dynamics of online news, mainly in tasks such as entity and topic extraction, and for visualizing the evolution of topic and entity coverage, as well as publication patterns over time. For instance, NLP has been used in identifying entities [2] and topics [7] in online news articles. Also notable is the connection with Human-Computer Interaction (HCI), which has been leveraged in creating visualization and interaction tools that support the exploration and interpretation of news data through interactive visualizations and user-friendly interfaces [25]. This last contribution, which is the MediaViz platform, focused on two main things: publication patterns and variations in coverage. Concerning the former, MediaViz enables the comparison of publication patterns in terms of the count and percentage of articles published by multiple sources according to several temporal granularities: weekly, monthly, and 24-hour cycles. Regarding the latter, the tool is also

intended to allow users to observe how different types of news sources cover various topics over time. By entering multiple search terms, users can track the number of articles — either in absolute terms or as a percentage of all articles published on a given day — that include those keywords throughout the selected time period.

Considering all the compiled knowledge, the importance of diving into the study of the online news landscape in Portugal is clear, mainly in what concerns variations in coverage and publication patterns. Before anything else, there is limited knowledge on the topic, as efforts leveraging different disciplines of informatics to study the Portuguese online news ecosystem are relatively few. Thus, it is a topic that has considerable potential for exploration and is of interest to many people. Firstly, media professionals, as this can contribute to a deeper understanding of digital journalism. For example, understanding trends and shifts in topic and entity coverage can help media professionals and news outlets understand which topics and entities are receiving more attention in a certain time interval, which can inform media practices. Finally, a better understanding of specific publication patterns can permit news outlets to evaluate their and their competitors' publishing behavior in a more effective way.

1.2 Problem

This proposal aims to study the Portuguese online news media landscape by characterizing a large collection of news headlines and summaries that is already prepared. We can consider two main goals here. The first goal is to describe the state of the art, by examining what has been done so far in this field. Although little work about the Portuguese news media landscape is known, there has been considerable research on this topic done in other countries, and, by inspecting what has been done in other contexts, we can better comprehend the path we need to follow in order to conduct a study such as this in the Portuguese context.

The second goal is related to the characterization of the dataset that contains news headlines and summaries from various online sources in Portugal, mainly concerning publication patterns and variations in coverage [25]. First, this work involves examining the publication patterns of news articles. These publication patterns include investigating the frequency and timing of news articles releases across different news sources, how these vary throughout the month, week, or day, and how such temporal trends reflect the operational habits and editorial routines of each media outlet. Then, we need to develop methodologies for topic and entity extraction within the collected dataset. These methodologies involve topic modeling [1] and topic labeling [28] for topic extraction and NER [68] for entity extraction. For both of these dimensions, we need to study the use of different visualization approaches to understand which are the best suited to comprehend the evolution of both publication patterns and variations in coverage over time. In addition to this, another dimension that can be explored concerns topic and entity diversity per media outlet.

1.3 Research Questions

So that we can address this problem, we can translate the aforementioned objectives into the following questions, which this work aims to answer:

RQ1: What is the best methodology to follow to extract topics from news documents?

RQ2: What is the best methodology to follow to extract entities from news documents?

RQ3: How can publication patterns of different news sources be analyzed through the volume and frequency of news articles published by them?

1.4 Document Structure

This dissertation includes five other chapters in addition to the introduction. This section summarizes the content of each chapter. Chapter 2 explains the methodology followed for the literature review on variations in coverage, publication patterns, and media diversity. Chapter 3 presents the state of the art, offering a comprehensive overview of current developments based on the relevant literature gathered during the literature review phase. Chapter 4 details the solution developed to meet the objectives described in the introduction related to topic and entity extraction and understanding of publication patterns. Chapter 5 presents the results obtained from the various experiments conducted, along with their corresponding analysis and discussion. Chapter 6 concludes the dissertation by summarizing the key findings and outlining potential directions for future research on the topic.

Chapter 2

Literature Review

In order to understand the current knowledge about the online news ecosystem in Portugal, mainly in relation to variations in coverage and publication patterns, a focused analysis of the research already done in this field is conducted here in the form of a literature review.

Throughout the years, many approaches have been designed in order to conduct studies in this field, as there has been a growing interest in understanding the dynamics and trends in news publication, which can be explained by the growth of information being published and disseminated nowadays. With this in mind, the aim here is to describe all the steps taken that resulted in the set of documents presented at the end of this section, which are aligned with the objectives represented in the research questions in Section 1.3. Figure 2.1 shows a flowchart that summarizes these steps, which are further discussed in the following sections.

2.1 Identification Phase

In this first phase, four digital libraries, namely IEEE Xplore[1], SCOPUS[2], ScienceDirect[3], and ACM Digital Library[4], were searched for relevant documents. To retrieve these documents, four search strings were created that summarize the main focus of this study. The use of search strings is crucial, as the results would be too broad if only individual keywords were used. Each search string went through various refinement phases so that the most relevant results possible could be retrieved. It is worth noting that both a search string concerning filter bubbles and another related to media bias were present in the initial iteration. However, as explained in Subsection 2.3.2, it was later decided not to delve into these two topics, and this decision led to the removal of these two search strings. However, these search strings are still shown here to document the preliminary work carried out, even though they no longer proceed to the refinement phase presented in Subsection 2.1.1.

[1] https://ieeexplore.ieee.org/Xplore/home.jsp
[2] https://www.scopus.com
[3] https://www.sciencedirect.com/
[4] https://dl.acm.org/

The first phase of search strings is shown below:

- **SS1**: "news" AND ("diversity" OR "plurality") AND NOT "fake"
- **SS2**: "news" AND "patterns" AND NOT ("consumption" OR "fake")
- **SS3**: "news" AND "visuali?ation" AND NOT "fake"
- **SS4**: "news bias" AND NOT "fake"
- **SS5**: "filter bubbles" AND NOT "echo chambers" AND NOT "fake"

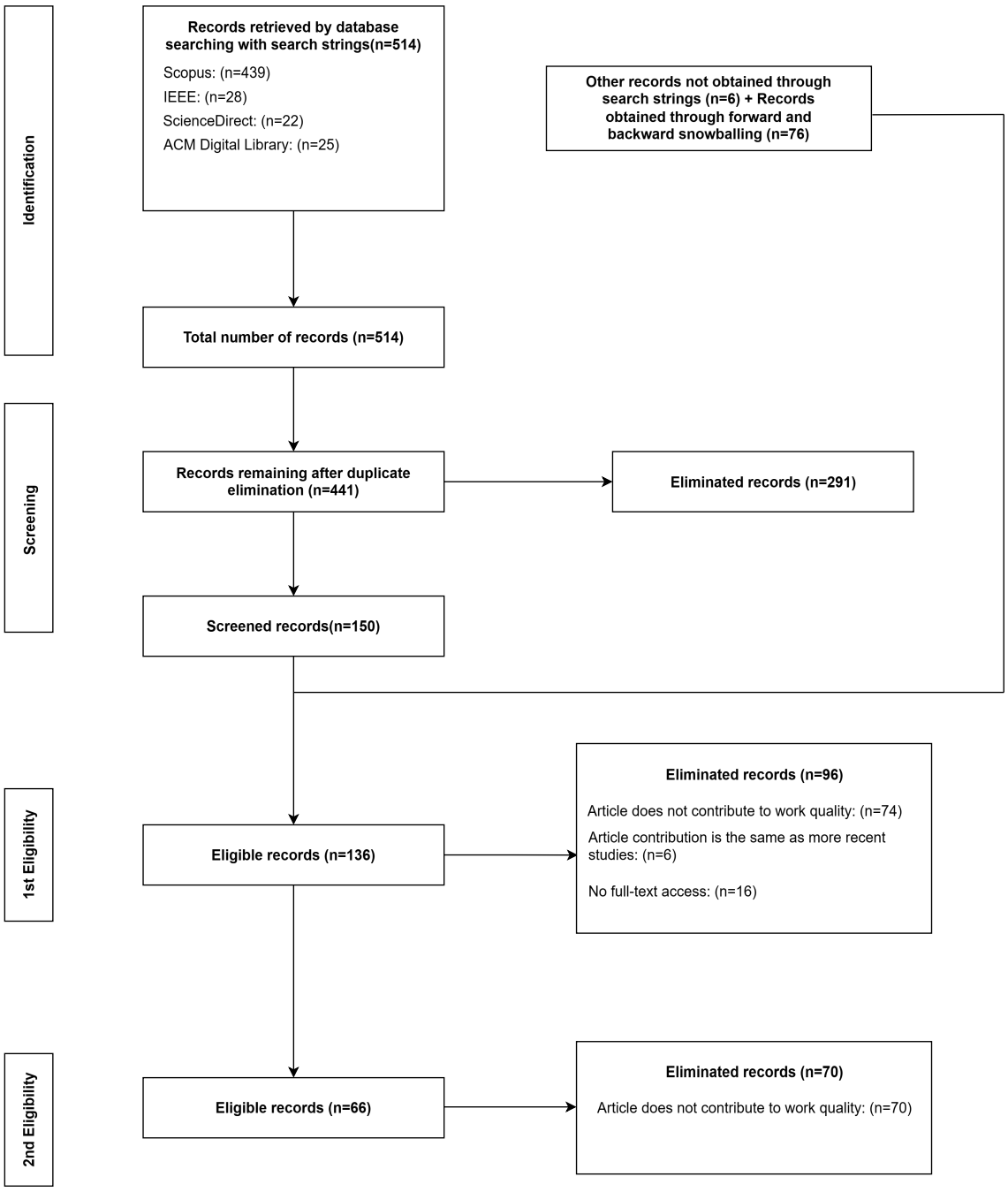

Figure 2.1: Literature Review steps flowchart

2.1.1 Refinement Phase

Taking into consideration that the first phase of search strings returned a total of 620 documents (711 if duplicates are taken into account) and that many of them were not closely related to the theme being studied, it was necessary to refine the search strings. The refinement was possible using the SCOPUS digital library as the main reference. Next, an explanation of the refinement steps for each search string is presented.

2.1.1.1 SS1 Refinement Explanation

Regarding this first search string, after analyzing some documents that were retrieved and understanding that there are other synonyms of "diversity" or "plurality" that are commonly used to represent the problem of news diversity, two synonyms were added to the query so that more related relevant results could be retrieved: "homogenization" and "commonality".

2.1.1.2 SS2 Refinement Explanation

In this case, many records not related to the theme being studied were being retrieved, as joining the keywords "news" and "patterns" led to too broad results. So, to mitigate this problem, the words "dynamics" and "dissemination" were also used alongside the word "patterns", as options that could be associated with the term "news". These words were chosen based on the recommended literature presented in the theme proposal. Also, a proximity operator was used to join "news" with the aforementioned words, instead of the AND operator, which caused the query to be too vague.

2.1.1.3 SS3 Refinement Explanation

Regarding this case, the term "data" was added in order to restrict the results more. This change led to more results about visualization of news publication or flow patterns.

2.1.2 Final Search Strings

After the changes made during the refinement phase, the following search strings emerged and represent the final set of queries used to retrieve documents:

- **SS1**: "news" AND ("diversity" OR "plurality" OR "homogeni?ation" OR "commonality")
- **SS2**: "news" NEAR/2 ("dynamics" OR "patterns" OR "dissemination") AND NOT ("consumption" OR "usage" OR "fake")
- **SS3**: "news" AND "data" AND "visuali?ation" AND NOT "fake"

The first and third search strings considered both the American and British English versions of the words "visualization" and "homogenization" by using the wildcard operator, so the results could contemplate document titles with both variations of these words. The motivation for the first

search string was to retrieve documents related to news diversity measurement in different contexts; for the second, the motivation was to get documents related to the study of news publication or flow patterns (some documents about variations in coverage were also retrieved); for the third query, the aim was to get documents that studied visualization approaches to assess the evolution of news publication or flow patterns. It is important to note that the word 'fake' was intentionally excluded from the search queries, as articles containing this term typically focus on fake news, which falls outside the scope of the present study.

These search strings resulted in a total of 514 retrieved documents. These were then exported as a BibTex reference and imported into the references management system being used (Zotero).

2.1.3 Search criteria

With the objective of restricting the number of documents being retrieved, some search criteria were defined:

- Selected records were restricted to journal articles, conference papers and book chapters;
- Only documents published in the English language;
- Only published records, i.e. , which were not work-in-progress documents.

2.2 Screening Phase

This phase encompasses two main steps: defining inclusion and exclusion criteria and screening the remaining records by analyzing their titles, abstracts, and conclusions in a detailed way to assess if they are relevant or not, according to the inclusion and exclusion criteria defined.

2.2.1 Inclusion and Exclusion Criteria

A total of 514 records were retrieved from the search strings. Given this, it is of extreme importance to define inclusion and exclusion criteria so that the most relevant documents can be selected out of those. The following criteria were defined:

- Article, conference paper or book chapter;
- Published in the English Language;
- Not work-in-progress publication;
- Not retracted publications;
- Not duplicated;
- Related to the study of dynamics of the news ecosystem;
- Related to the domain of news diversity measuring, publication or flow patterns;

- Related to the domain of variations in topic or entity coverage;
- Not related to the domain of fake news, rumors, news consumption/interaction with users, patterns of news not in textual format (images, videos, etc.);
- Available.

2.2.2 Removed duplicates

Duplicates were first identified and merged using the Zotero feature that finds duplicate documents. Then, in the case of some duplicates that could not be identified by Zotero, they were manually eliminated if their authors and titles fully or partially matched.

2.2.3 Screened records

After the elimination of duplicate records, a total of 441 documents were left. In this phase, each document's title, abstract, and conclusion were carefully analyzed to decide if the document was relevant or not based on the criteria detailed in Section 2.2.1. This process culminated in the elimination of 291 records, with 150 records left.

2.2.4 Forward and Backward Snowballing

The next step before starting the eligibility phase was to group more documents, but this time using different methodologies. First, backward and forward snowballing techniques were used. Backward snowballing consists of examining the bibliographies of publications so that more relevant results can be retrieved. Forward snowballing, on the other hand, consists of analyzing the documents that cite the selected records, in order to get more recent results. This resulted in 76 new relevant documents. After this, 6 more documents were collected through fortuitous searches that were made before the search strings were defined. All these 82 records' titles, abstracts, and conclusions were previously analyzed and considered relevant, entering directly into the set of documents ready for the eligibility phase.

2.3 Eligibility Phase

2.3.1 First Eligibility Phase

The eligibility phase starts with 232 documents (the aforementioned 82, and the 150 from the screening phase). Although a complete full-text analysis of these records was not performed, each of these was further analyzed to restrict the final number of selected papers even more. Concretely, in addition to a complete skim of each article, the results section of each paper was carefully analyzed to determine whether the study addressed the same research objectives pursued in this work. This resulted in the elimination of 96 papers, with 136 records left. The reasons for the

elimination can be seen in Figure 2.1. Article does not contribute to work quality" refers to a misalignment between the article's focus and the objectives of the present research.

2.3.2 Second Eligibility Phase

In the second eligibility phase, which involved a full-text review of every paper, 70 of the 136 papers from the previous eligibility phase were excluded, remaining 66 papers. Notably, all papers related to the filter bubbles and media bias topics were among those removed. In what concerns the filter bubbles topic, this decision was made because there is still little consensus regarding this topic, as there are studies that discredit the existence of filter bubbles [21], while there are others that defend its existence [53]. Additionally, assessing filter bubbles requires measuring exposure diversity, which depends on news consumption data, which is information that we do not have access to. Regarding media bias, the exclusion was motivated by the fact that labeling a news article as biased or not can be reductive. Bias is a complex phenomenon that cannot always be clearly identified at the article level. It may depend on broader contextual factors such as framing, source selection, or comparative analysis across multiple outlets or time periods. As such, a simplistic classification of bias might not capture the nuanced ways in which media bias manifests.

In Table 2.1 the articles that are not mentioned in Chapter 3 and that do not belong to the filters bubble or the media bias topic were still kept in the tables to show the investigation work that was carried out. In order to distinguish them from the ones that are mentioned in the state of the art, they are marked with a "*".

Table 2.1: Selected records after the second eligibility phase

Theme	References
Variations in Coverage	[22], [25], [63], [24], [55], [94], [30], [26], [46], [79]*, [95]*, [83]*, [29], [45], [9], [20], [47]*, [34]*, [15]*, [73]*
Publication Patterns	[51]*, [50]*, [56]*, [97]*, [52]*, [93]*, [99]*, [13], [23]*, [27]*, [6], [80]*, [14], [36], [18], [31]
Media Diversity	[92]*, [44]*, [4], [33], [11]*, [8]*, [39]*, [19]*, [17]*, [72]*, [57], [29], [59], [76]*, [42], [78], [89], [41], [61]*, [90], [66], [100]*, [69], [82], [91]*, [88], [54], [62], [37], [3]

2.4 Summary

The literature review described in this chapter aimed to identify and analyze relevant research that informs the three main pillars of this dissertation: variations in coverage, publication patterns, and media diversity.

To accomplish this, multiple phases were needed. First, a comprehensive search strategy was executed using four major digital libraries, including IEEE Xplore, SCOPUS, ScienceDirect, and ACM Digital Library. Then, five initial search strings were created, though two (on filter bubbles and media bias) were later discarded due to conceptual complexity and lack of necessary data, and the search strings that remained were refined using proximity operators and synonyms to improve relevance. Finally, a multi-phase filtering process (screening, eligibility, and full-text review) narrowed the initial 711 records down to 66 selected studies. These final references were distributed across the three main themes: 20 papers on variations in coverage, 16 on publication patterns, and 30 on media diversity.

In total, the chapter lays a solid empirical and methodological foundation for the state-of-the-art analysis that follows in Chapter 3. It also emphasizes a careful curation of literature to ensure that all reviewed works are directly relevant to the dissertation's research questions and goals.

Chapter 3

State of the Art

In this chapter, a broad view of the state of the art is presented. This chapter is divided into three main topics related to the problem being studied: variations in coverage, addressed in Section 3.1, news publication patterns, approached in Section 3.2 and media diversity, tackled in Section 3.3. The presented state of the art comes as a result of the literature review presented in Chapter 2.

3.1 Variations in Coverage

Regarding variations in coverage, this theme is examined from two complementary perspectives: topic coverage and entity coverage. Topic coverage refers to the prominence of various topics addressed by news articles over time, while entity coverage focuses on the presence and visibility of specific named entities such as people, organizations, and locations within the news. To meaningfully assess these dimensions, it is essential to apply robust techniques for extracting both topics and entities from textual data. The former involves the utilization of topic modeling techniques, while the latter involves the use of NER.

3.1.1 Topic Coverage

Topic coverage is related to the attention given by news publishers to a certain topic or event. Understanding it and its evolution over time can be relevant to comprehend how certain issues gain or lose prominence and how editorial focus shifts in response to societal, political, or economic developments. First, topic modeling is assessed by analyzing the different techniques that are used to approach this problem. Then, topic labeling is also assessed. Finally, the focus shifts toward different efforts that approach the problem of the attention given to topics over time.

3.1.1.1 Topic Modeling

Topic modeling has the main objective of organizing massive collections of documents, as it works as a dimensionality reduction technique by mapping the documents in a topic space and clustering them based on their proximity to the detected topics [55]. Several topic modeling techniques have

already been used in studies on topic coverage [63, 64]: LDA, hierarchical Latent Dirichlet Allocation (hLDA), Structural Topic Model (STM), Dynamic Topic Model (DTM), discrete Dynamic Topic Model (dDTM) and k-means, to name a few. Here, the focus shifts towards the two studies in which all these techniques were used and compared.

Mele et al. developed an effort in topic modeling that compared the results obtained by four different techniques, namely hLDA, dDTM-News — a variation of dDTM — tailored to deal with highly dynamic streams of news documents [63], k-means and k-means+time, which is based on k-means with the twist that it eliminates news documents that are outside the time window of the event of interest [63]. First, news documents were labeled at the level of fine-granularity events to evaluate the performance of the clustering. Then, for each of the 60 events that resulted from the labeling, news documents from Wikipedia that had a high cosine similarity to the event keywords were collected. This collection process was evaluated by crowdsourcing, resulting in 4300 pairs event, news and corresponded to the phase of data creation. Then, event detection was performed and DTM, dDTM-News and two other techniques, namely Unigram Co-Occurrences [94] and Event Detection with Clustering of Waveletbased signals (EDCoW) [30] were compared, being dDTM-News the one which performed the best in terms of event coverage, being able to cover 55 out of the 60 events. Finally, dDTM-News was compared to k-means, k-means+time, and hLDA for news clustering. This was measured as the proportion of documents relevant to a specific cluster relative to the total number of documents in that cluster, and again, dDTM-News proved to be the best for this task, reaching an F1-score of 0.83. Concerning the approach taken by Mendez et al. on topic coverage [64], they used STM to allow the incorporation of contextual variables during the topic modeling process. In the case of this study, the main contextual variable was the territorial level of the news source. The Subsection 3.1.1.3 explores this process in greater detail.

3.1.1.2 Topic labeling

Topic labeling involves assigning labels to topics generated by a topic modeling algorithm. This step is crucial, as the representations produced by topic modeling tools are not always easily interpretable by humans [28]. With regard to this, there have been two different approaches to the problem of assigning labels to topics.

The first strategy concerns the use of an external knowledge base, namely GermaNet [28]. The authors propose an automatic topic labeling method that combines topic modeling with word sense disambiguation using this knowledge base. After applying the Top2Vec algorithm to extract topics, they disambiguate the senses of ambiguous topic words based on their semantic similarity to unambiguous words in the same topic. Two similarity measures are used: Jiang-Conrath similarity [43] and a vector-based similarity using pre-trained German word embeddings, hypernyms, and hyponyms. The goal is to determine the most contextually appropriate sense for each ambiguous word, enhancing the interpretability of the topic label.

The second study also leverages an external knowledge base, this time English Wikipedia[1],

[1] https://en.wikipedia.org/wiki/Main_Page

to automatically label topics derived from LDA [48]. First, it generates label candidates using three sources: Wikipedia article titles retrieved by querying with the top topic terms; sub-phrases extracted from these titles via chunk parsing; and the top five topic terms themselves. Candidates are then ranked using a combination of lexical association measures (e.g., Dice, t-test), lexical features (e.g., term counts, overlap with topic terms), and an information retrieval score based on Wikipedia searches. These features are fed into a supervised support vector regression model, trained on human-annotated data from Amazon Mechanical Turk [2], to select the best label per topic.

3.1.1.3 Topic Attention Over Time

After talking about topic modeling and topic labeling, it is now time to delve into the specific problem of topic coverage and consider the several approaches that have been developed concerning it. In the paper by Devezas et al. , a visualization platform for online media studies called "MediaViz" was described [26]. It encompasses two components: the first is a back-end application that fetches and stores data from news articles from various online news sources via their Really Simple Syndication (RSS) feeds and provides access to them through an Application Programming Interface (API), while the second is a client-side application that retrieves those data and allows their exploration through a series of interactive visualization tools. One of them, the "Keywords tool" [25], showed the coverage that each type of source considered in the study, namely news sources and blog sources, gave to different search terms over time, as shown in Figure 3.1 and in Figure 3.2, respectively.

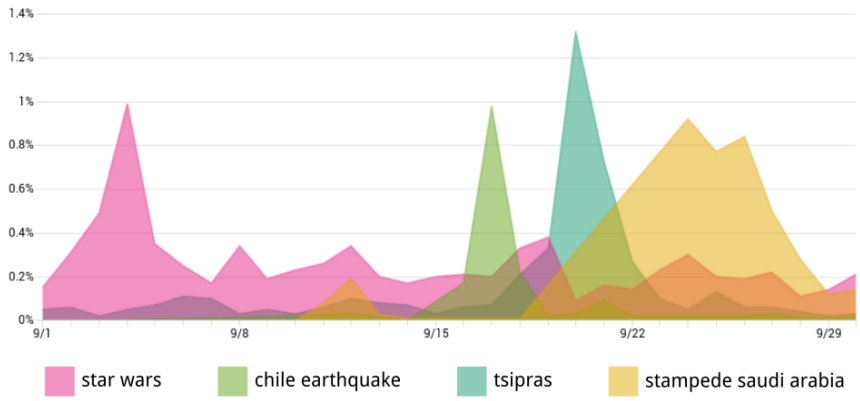

Figure 3.1: Visualization example of the "Keywords tool", showing the percentage of articles published about each of the topics in news sources [25].

[2] https://www.mturk.com/

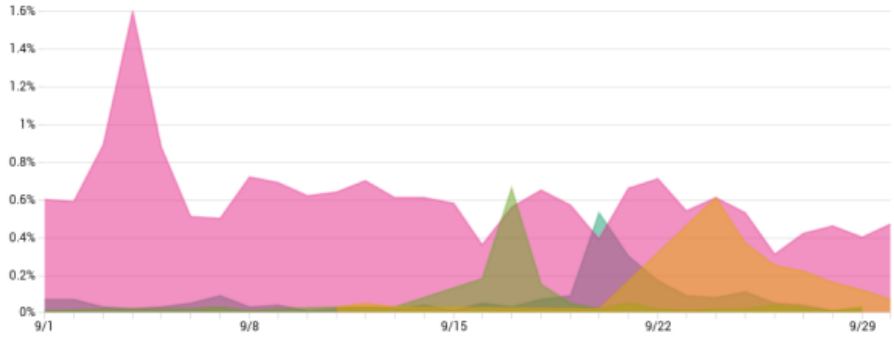

Figure 3.2: Visualization example of the "Keywords tool", showing the percentage of articles published about each of the topics in blog sources [25].

With a similar objective in mind, Koivunen-Niemi and Masoodian also developed a visualization tool called "time-sets" consisting of a modified form of linear diagrams that made it possible to examine how topic choices evolved chronologically between different news sources [46]. A specific topic or keyword was considered a set to which the articles that contained that topic or keyword were appended. Each set would then be compared to the others to contrast the attention that was being given to each topic. This was achieved by first collecting a series of news articles from different news sources through scraping and then storing them in a database. After that, the articles were processed with Python natural language processing tools to detect the presence of each keyword or topic in the article. If the topic appeared in it, the article would be assigned to the set. Finally, Flask[3] was used to filter and feed data to the "time-sets" visualization tool, which was developed using NodeBox Live[4]. A visualization example of this tool can be seen in Figure 3.3.

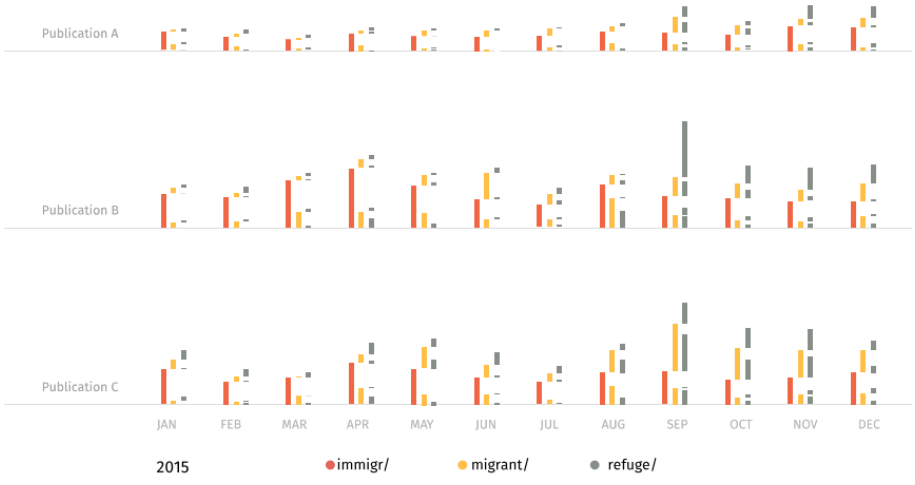

Figure 3.3: Time-sets visualization showing the amount of attention given to each of the topics mentioned across three different news publication sources [46].

[3] https://flask.palletsprojects.com/en/stable/
[4] https://nodebox.live/

Finally, three studies that took a slightly distinct approach that also assess the theme of coverage are those by de León and Vermeer, by Mendez et al. , and by Devezas et al. [22, 64, 25]. One of these studies' main objectives was to evaluate how topic choices changed under specific periods or circumstances. The paper by de León and Vermeer [22] took an approach based on distant supervised machine learning to classify articles as "political" and "non-political" based on the Uniform Resource Locator (URL) content of the news articles. Then, a set of 16 classifiers, which corresponded to two classifiers by country in two different periods, one in a regular period and the other during an election period, were trained with the Naive Bayes algorithm and tested against the corpora of news articles to classify them according to the two aforementioned labels. The quantity of political news published during the election period increased for most countries, although not astonishingly, as shown in Figure 3.4. Nevertheless, it should be noted that the results are inevitably affected by the classification of news into "political" or "non-political" which is not perfect.

Figure 3.4: Difference in percentage of political articles posted between election and routine periods [22].

Then, the paper by Mendez et al. proposed a topic modeling procedure to attack the problem [64], using the methods described above in this subsection. This paper, which explores the territorial and temporal patterns of European Union (EU) cohesion policy media coverage by using topic modeling, also proves that topic choices vary a lot according to different circumstances, as it allowed the authors to conclude that topic choices differ at different geographical levels, i.e. , topics that may be more addressed by regional news sources may not receive the same attention at the national level, as shown in Figure 3.5. In this situation, STM was valuable, as it allowed the authors to comprehend the differences between the attention given to different topics depending on the geographical context. In addition, it is also shown that topic prevalence varies between

countries over time, as some topics may be relevant at different times in different parts of the world, as presented in Figure 3.6.

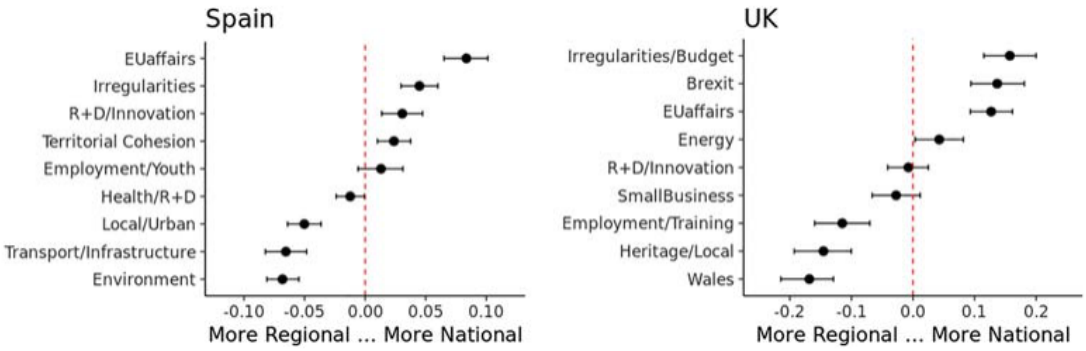

Figure 3.5: Topic coverage between different geographical levels [64].

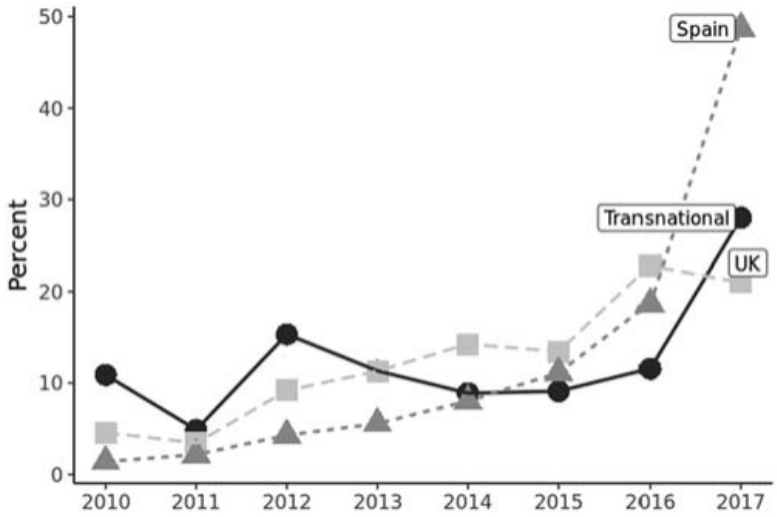

Figure 3.6: Coverage on cohesion policy over time in different geographical contexts [64].

To conclude, the paper developed by Devezas et al. introduced yet another visualization tool called "NewsMap" which presents a choropleth world or Portuguese map divided by administrative units [26]. This visualization tool can show a source or group of sources' geographic coverage or even the amount of coverage a certain geographic region received in a user-chosen topic. In addition, the tool also offers the possibility to limit the search to a specific period. A visualization example of this tool can be seen in Figure 3.7.

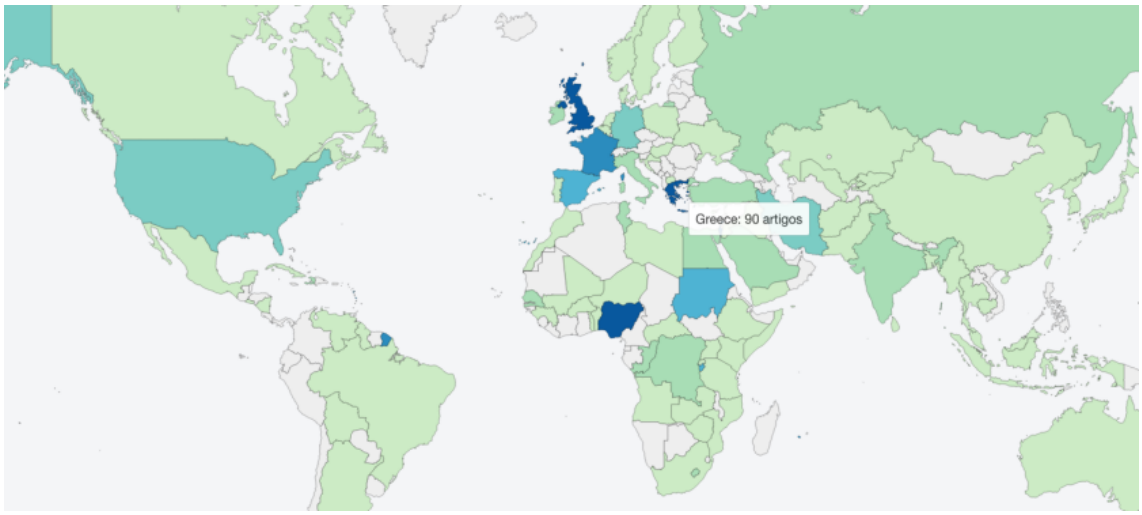

Figure 3.7: Visualization example of the "NewsMap tool", showing the coverage the term "election" received worldwide [26].

3.1.2 Entity Coverage

Entity coverage refers to the attention that news publishers allocate to specific entities, whether they are people, locations, organizations, or other types of named entities. Comprehending entity coverage and its evolution over time can provide valuable insights into how the visibility of specific individuals, organizations, or places changes in the media landscape.

Several studies have delved into the detection of named entities. Kiritoshi and Ma [45] utilized StanfordCoreNLP[5], more specifically its Stanford's Named Entity (NE) recognizer tool to extract named entities from news documents. Another effort that extracted named entities from tweets using Stanford's NE recognizer tool was the one developed by Elejalde et al. [29]. Another example, particularly dedicated to the identification of geographical entities in text, is presented in the paper by Chen et al. [20]. Here, after translating a news text in any language to English, using machine translation tools, such as Google Translate or Microsoft Translator, an alignment algorithm is used to map words and phrases in the source language to their counterparts in the English translation. Finally, the translated English text is processed using an English geoparser, which identifies and extracts location names, and the identified locations are matched back to their corresponding terms in the original language using alignment data. The output of this process consists of the names of the identified locations, provided in both English and the source language.

Neural models have also been used in entity detection. An example of these neural models is a model called Generalist Model for Named Entity Recognition using Bidirectional Transformer (GLiNER) [98]. GLiNER is an open-domain NER model designed to extract entities of any type using a lightweight and efficient architecture. Unlike traditional NER models limited to predefined categories, GLiNER allows users to specify custom entity types via natural language prompts. It leverages a bidirectional transformer encoder to jointly encode both the input text and the entity

[5] https://stanfordnlp.github.io/CoreNLP/

type descriptions. The model computes embeddings for all possible spans in the text and matches them with entity-type embeddings in a shared latent space. Using a span-based classification approach, GLiNER can identify multiple entity types in parallel, offering fast and accurate predictions. This makes it especially suited for scalable and resource-constrained applications. GLiNER demonstrates strong performance, outperforming both ChatGPT and fine-tuned LLMs in zero-shot evaluations on various NER benchmarks [98].

3.2 News Publication Patterns

This section presents an overview of news publication patterns, which play a key role in understanding the temporal dynamics and operational routines of different media outlets. These patterns provide insights into how frequently news content is produced, the typical times and days of publication, and how these aspects vary between sources. Such an analysis can reveal structural differences between outlets, uncover editorial rhythms, and highlight disparities in content volume over time. A more detailed discussion of publication and update patterns is provided in Subsection 3.2.1.

3.2.1 Publication and Update Patterns

In this subsection, the focus shifts towards publication and update patterns. Comprehending the publication and content update characteristics that define different news outlets can be important for several reasons. First, as emphasized by Avraam et al. , who made a study about publication patterns on Greek news websites [6], it can be important for news outlets because this understanding can allow them to develop effective content publication routines, which can contribute to the optimization of journalistic practices. In a slightly different but also relevant perspective, Calzarossa and Tessera stated that search engines must keep up to date with all the content published or updated on news websites and have a timely way to refresh and index this content to make it valuable to users [13]. Therefore, it is crucial to predict how the content of these extremely dynamic websites changes so that search engines can adjust their crawling activities accordingly [13].

Devezas et al. presented a visualization tool that studied the publication patterns between two different source types, namely news sources and blog sources, called the "Sources tool" which is capable of comparing the publication patterns between multiple sources at different time granularities: daily, weekly, and monthly [25]. In Figure 3.8, a comparison between news and blog sources in terms of the percentage of articles published by each of them in the temporal space of a week is presented.

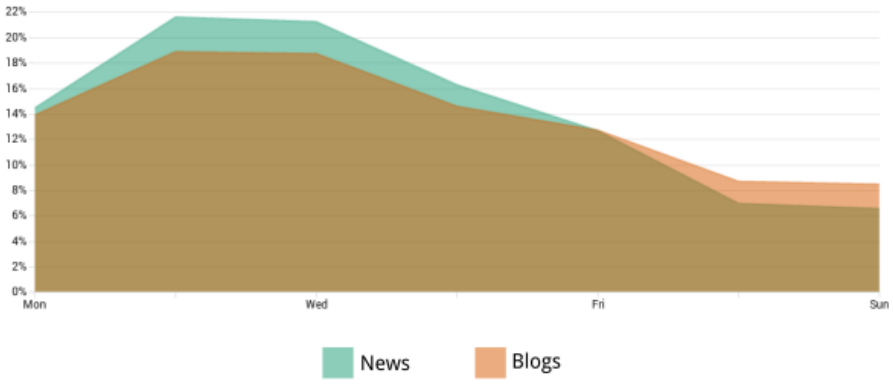

Figure 3.8: Example of visualization of the "Sources tool", showing the percentage of articles published by each of the sources [25].

Going back to Calzarossa and Tessera, they developed an extension of their previous paper [13] with the same motivation in mind. The objective was to study the properties and temporal patterns of the change rates of the content of three news websites and predict the future temporal patterns of the change rates of the content of the websites based on past temporal patterns [14]. To understand these temporal patterns, they used time series analysis [36], and here they divided them into trend, seasonal, and irregular components. The trend component corresponds to long-term variations, while the seasonal to periodic patterns (e.g., daily, weekly). They are modeled using trigonometric polynomials. The irregular is associated with residual, non-structured variations and it is represented by an Autoregressive Moving Average (ARMA) model [18]. The overall time series is obtained by adding the three components. To forecast the future dynamics of the web content, the predicted value of the time series Y^t+h at time t+h is obtained by superimposing the values predicted by these models. For the trend and seasonal components, the values are extrapolated from the corresponding models computed at time t+h, while a moving horizon prediction technique based on the Box-Jenkins approach [31] is applied to forecast the values of the irregular component. To observe the content change rates of each website, a set of pages and the number of changes made to them were crawled and plots were created to show the temporal patterns of the content changes, as shown in Figure 3.9. The predictions of future temporal patterns of the change rates of the websites' content were validated by comparing predictions against unseen empirical data over different time horizons (1, 3, and 6 hours), as shown in Figure 3.10. Descriptive measures, such as absolute and relative errors, were used for evaluation. This exploration allowed the authors to identify daily and weekly patterns in content change rates and accurately forecast the web content's future dynamics for short time horizons.

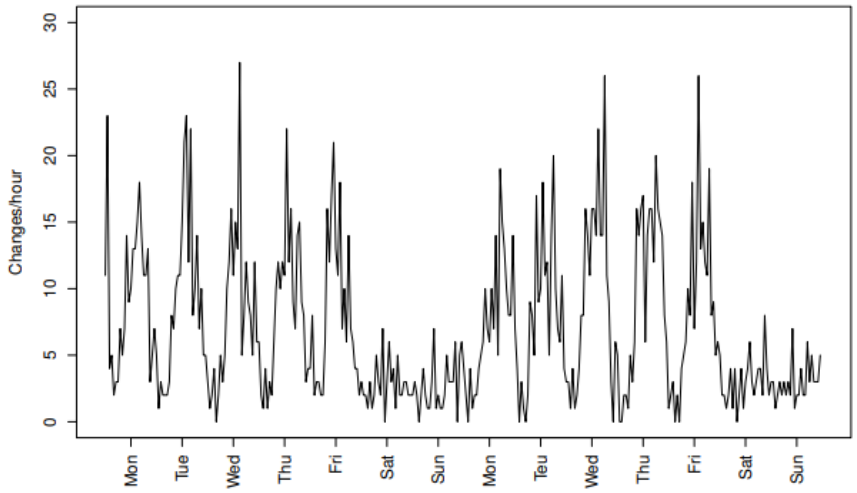

Figure 3.9: Temporal patterns of content change rate for the MSNBC news webiste [14].

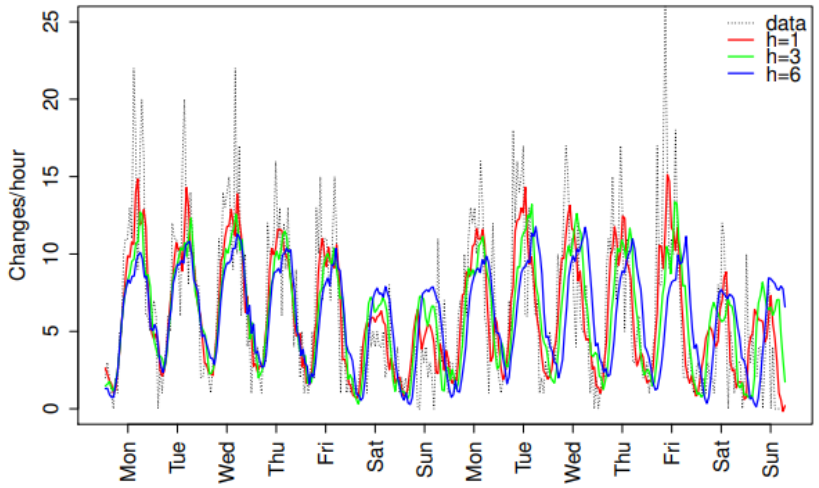

Figure 3.10: Prediction of the change patterns over three different time horizons (1, 3, and 6 hours) [14].

In a different perspective, Mele et al. proposed an approach whose purpose was to measure the timeliness score of a news stream, which is defined as a sequence of news published by a newswire using a publishing platform [63]. Timeliness calculates, for each event reported in a news stream, the complement of the average time delay of the stream in reporting the event normalized to a scale of [0,1]. The delay can be characterized in two ways: the first considers the number of streams that preceded the stream in reporting the event, while the second considers the time distance between the considered stream and the first stream that published about the event. In this way, different news streams can be ranked according to the latency with which they report a specific event.

3.3 Media Diversity

Following the analysis of variations in coverage and publication patterns, a topic about diversity is now presented. First and foremost, a general overview is introduced in Subsection 3.3.1, which focuses on diversity on a conceptual level to give a clearer view of what media diversity consists of. Then, the attention turns to diversity on an empirical level, first with the introduction of different frameworks that are sometimes used in the evaluation of media diversity in Subsection 3.3.2 and then with the introduction of various techniques that have been used to measure diversity in Subsection 3.3.3.

3.3.1 General Overview

Diversity is considered to be of crucial value both for communications and media policy [88]. Many definitions concerning the concept of diversity have been presented [90, 66, 62], each of them approaching the concept from different perspectives and using different dimensions to categorize it. It is important to note that there are many more definitions than the ones presented. Many studies approach diversity on a more empirical level, with the purpose of measuring diversity, and thus define the concept by the procedure of the measurement they use, through a formula or a statement [88]. This case can be observed in the paper by An et al. , where diversity was defined by the following statement: "We compute the political diversity of news articles [...], which is defined as the (political) entropy of the news articles associated with the user" [5]. Nevertheless, the ones that are detailed come from central studies in this field and are the ones that are still cited as the main theoretical basis for the characterization of the concept of diversity in the media field.

At the highest level, Voakes et al. [90] do not address anything but the media content, specifically focusing on diversity in terms of who is mentioned, namely people, groups, or organizations, and what perspectives are taken. These two views of diversity encompass two sub-dimensions of media diversity, namely entity and viewpoint diversity [54]. On the next level, Napoli [66] adds on what Voakes had previously proposed and includes another element to content diversity evaluation, this time related to the events that are covered. This view of diversity includes a sub-dimension of diversity called topic diversity [54]. In addition to adding a topic element to the concept of diversity introduced by Voakes et al. , Napoli mentions the structural component, which is related to who made the content and how it was made, and this view was coined as source diversity (it can also be classified as structural diversity, as labeled by Loecherbach et al. [54]). He adds exposure diversity as another component to the definition, but it is not further explored in this section, as the focus here is the diversity in news production. Finally, McQuail [62] provides the most detailed account of diversity, and, besides talking about content, structural, and exposure diversity, he also talks about different goals that can be addressed with diversity as, for example, reflecting the differences of a population (reflection), and giving the opportunity of separate groups of society to speak to the wider society (access). McQuail also introduces performance standards to evaluate diversity, such as equality and proportionality. Nonetheless, more recent studies that elaborate

further on this idea by refining the performance measures introduced by McQuail are detailed in the following section.

3.3.2 Frameworks for Media Diversity

To answer a question as broad as "How much diversity is enough?" it is crucial to establish different benchmarks for evaluating media diversity, as this question can be answered differently depending on the objective of the evaluation. Four distinct normative frameworks, each with its focus, key goals, and benchmarks for measuring diversity have been proposed: liberal-aggregative, liberal-individual, deliberative, and adversarial [12, 38, 65, 71]. The first model is the most popular among highly cited scholars, such as Napoli and Voakes [54], and focuses on reflecting social heterogeneity, as the benchmark for diversity consists in mirroring society, adopting a perspective of reflective diversity [88], i.e. , the diversity portrayed in the news should match the diversity observed in society. The deliberative framework focuses on promoting inclusive public debate to discover all possible perspectives, and the benchmark for diversity consists in promoting equal shares of viewpoints and entities and inclusion of all actors in the debate, adopting the perspective of open diversity [88]. The adversarial model has the main objective of giving voice to minorities, and its benchmark for diversity involves supporting the share of minority voices and small groups on the news. The liberal-individual prioritizes consumer satisfaction, and it is not addressed here, as it focuses on perceived diversity. Nevertheless, few papers provide a comprehensive discussion of normative frameworks [89, 57].

Before delving into the process of media diversity evaluation, it is also important to turn our attention to the framework built by Stirling, who coined the terms "variety", "balance", and "disparity" to describe distinct approaches used to measure diversity [82]. The first can be described as the number of different elements present in a news document. For example, when analyzing a news article in terms of topic diversity, the news article is considered to be more varied and thus more diverse the more topics it presents. However, this is not a complete way of measuring diversity, as it does not capture the coverage given to each of those topics. The second corresponds to how much of each element there is. In the case of topic diversity, the more distributed the attention given to each topic is, the more balanced and diverse the news document is. Finally, the last considers the difference between the elements being highlighted. Returning to the example of evaluating topic diversity, the more distinct the topics presented in a news article are, the more diverse it is.

The following section presents a set of empirical-level studies on the evaluation of media diversity.

3.3.3 Media Diversity Evaluation

The three sub-measures of diversity presented previously in Subsection 3.3.2, namely variety, balance, and disparity, are fundamental concepts in media diversity evaluation [82]. In fact, Elejalde et al. refuted Napoli's assumption that a high number of different available news sources may lead

to a wide variety of different topics, actors, and opinions within single media outlets [66], i.e. , that high variety by itself leads to high diversity. They did this by conducting a study of the ecological diversity and health of online news in Chile, where they found that this country, although having many news outlets responsible for the publication of news, i.e. , high structural diversity, they were all controlled by few owners, which were proven to influence the outlets' editorial policies. This resulted in low levels of diversity in the news [29]. Sjøvaag and Pedersen [78] utilized these three distinct metrics to analyze the differences in topic diversity between Norwegian newspapers with and without direct press support. LDA was run using the MALLET toolkit [60]. The model was tested with several topic numbers and the one with 200 proved to be the best, although 14 of them were deemed as incoherent and thus were not selected, remaining 186 topics. A topic is represented by a cluster of frequently co-occurring words, and a document is about the topic with the largest weight and about every topic with a weight bigger than or equal to 50% of the weight of the leading topic. In terms of variety and balance, no differences were observed: every topic was present in both corpora and topic distributions mirrored each other. In terms of disparity, it was evaluated in terms of unique contributions by press-supported newspapers. Here, the discrepancies were a little bit more visible, as the differences between the topics covered by the two corpora are more notable. Although this study led to relevant conclusions, LDA poses the limitation of assuming a fixed number of topics, potentially missing finer-grained topic variations.

3.3.3.1 Classical Methods for Measuring Diversity

The two most common formulas used to measure diversity in terms of variety and balance [82] are Shannon's H or its standardized version [75] and Simpson's D or Simpson's Dz, its standardized version [77], as they have been used in various studies on media diversity evaluation [33, 4, 42, 41, 59]. Another formula is the Herfindahl-Hirschman Indices (HHIn) [3], which has also been applied to measure diversity and compared to Shannon's H [57].

Amsalem, Guo, and Humprecth and Esser [4, 33, 42] used Shannon's H standardized version to measure different sub-dimensions of diversity. In the formula, p_i is the proportion in the ith category where categories = i through j, while k is the number of categories:

$$H = -\frac{\sum p_i \log_2 p_i}{\log_2 k}$$

Amsalem developed a computational framework to compare the diversity of topics between elite and popular newspapers, which involved text classification and measuring diversity [4]. After using a Deep Learning (DL) model to identify topics in news articles, the salience of each topic in the article was calculated and the standardized version of Shannon's H was used to calculate the diversity (Table 3.1). One notorious limitation to emphasize here is that automated methods of topic identification, such as the one used here, although being more time efficient, risk repeating classification errors, which gives an advantage to manual methods performed by human coders, who can learn and improve.

Table 3.1: Topic saliency and Shannon's H result for each category [4]

	Category 1	Category 2	Category 3	Entropy
Article 1	100%	0%	0%	0
Article 2	90%	10%	0%	0.3
Article 3	50%	50%	0%	0.63
Article 4	33%	33%	33%	1

Guo developed an approach to study Media Agenda Diversity (MAD) within China's controlled media environment, more specifically the differences between official and commercial news websites. It first used LDA to identify 31 topics in the news corpora being evaluated. After this, an SVM model was trained with 2050 manually labeled articles, and the model reached acceptable performance in identifying 26 out of those 31 topics in the training set, and was thus used to identify those 26 in the testing set. The MAD index [69] was used to group those topics into 15 themes, which were used to calculate topic diversity using Shannon's H. Humprecht and Esser measured entity and geographic diversity in six countries using this formula too. Entity diversity was focused on six elite actors and geographic diversity examined the representation of five different geographical regions in news documents. Finally, a one-way independent Analysis of Variance (ANOVA) was used to identify the influence of ownership types in news diversity. In both study's cases, the main limitation was the small temporal window from which articles were collected.

Although Shannon's H is the more commonly used measure, Magin et al. presented two advantages of using HHIn [57] — which, instead of measuring the presence of diversity, measures the absence of concentration — instead of Shannon's index. First, the HHIn offers a distinct advantage with its fixed value ranges, which make interpretation easier. Specifically, values below 0.15 indicate low concentration and diverse reporting, values between 0.15 and 0.25 reflect moderate concentration and values above 0.25 signify high concentration and less diverse reporting. Additionally, unlike Shannon's H, the normalized HHIn provides a measure that can be directly compared to other concentration metrics, even when the number of topic or actor codes varies from those used in the study in question. This was one of the papers that made reference to the normative frameworks, as indicators for diversity were derived from liberal and deliberative democratic theories. An example of this is the indicators derived to calculate entity diversity, as the ones derived from the liberal model, which focuses on reflective diversity, concern with evaluating the proportion between executive and legislative actors, while the ones derived from the deliberative model, which focuses on open diversity, concern with evaluating the proportion between executive and civil society actors, as the main focus of this model is to promote inclusive public debate, instead of giving voice only to the already most powerful entities in society.

Other studies opted to use different formulas to measure diversity, namely Simpson's Dz, which was used by Humprecht and Büchel and Masini et al. [41, 59]. In the formula, p_i^2 is the

proportion in the ith category where categories = i through j, while k is the number of categories:

$$D_z = \frac{1 - \sum p_i^2}{1 - \frac{1}{k}}$$

In the case of Humprecht and Büchel, articles were manually coded for topics and entities (actors) and in the case of Masini et al. , they were coded for entities (actors) and viewpoints. Then, the standardized version of Simpson's D was used to calculate the corresponding diversities in each case. An interesting approach taken by these studies was the exploration of factors that could potentially affect diversity. Humprecht and Büchel analyzed the influence of organizational factors, such as the resources available to outlets, and national-level factors, such as unemployment rates and national debt, on the diversity of the coverage of the "Occupy Wall Street" social movement. To do this, they employed a method called Qualitative Comparative Analysis (QCA), which allows for identifying combinations of national and organizational conditions conducive to high or low diversity. In turn, Masini et al. also used ANOVA to develop multilevel regression models to test the impact of article features, such as article length, and of newspaper characteristics, such as geographical reach.

In the next section, the focus shifts towards studies that did not leverage these classical approaches. A set of alternative methods to measure media diversity is described.

3.3.3.2 Alternative Methods for Measuring Diversity

Redirecting our attention to disparity, there have been fewer studies that have developed approaches that can be used in its evaluation. An example is the study developed by Devezas et al. [25], who also presented a visualization tool to observe the evolution of diversity over time. First, an element-wise aggregation was performed on the topic vectors for each day, selecting the maximum weight for each n-gram. This process produced a set of daily vectors that captured the topical direction for each day. After this, a combined distance metric was used to assess topic diversity, which allowed to compute the normalized cosine distance between all pairs of daily n-gram vectors, separately for each corpus, namely news and blog articles. A high mean cosine distance indicates greater disparity between topics, signifying higher diversity. Finally, a diversity score was obtained by combining the mean E[X] and standard deviation $\sigma(X)$ of the cosine distances.

$$score(X) = E[X] - 2 \times (F(E[X]; 0.5, 1/50) \times 0.5) \times (1 - E[X]) \times (E[X] \times \sigma(X))$$

$$F(x; \mu, s) = \frac{1}{1 + e^{-\frac{x-\mu}{s}}}$$

These diversity scores were then used to create a plot showing the evolution of diversity in both news and blog articles over the course of a month, as shown in Figure 3.11.

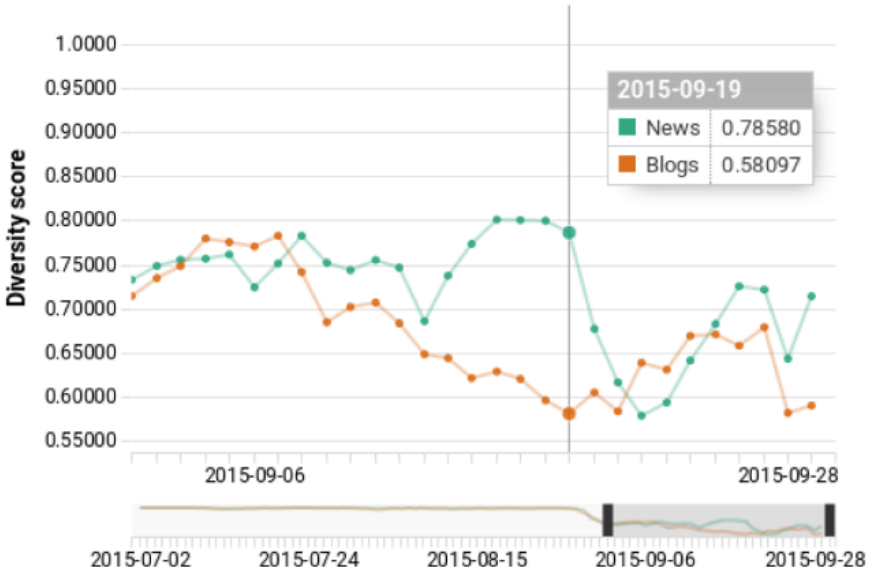

Figure 3.11: Diversity over time for windows of 5 days in the month of September [25].

Another distinct approach used a content analysis method called Network analysis of Evaluative Texts (NET) [87], which was used to encode political news into structured networks of actors and issues [89]. The objective was to investigate how media reflect political realities during election campaigns in the Netherlands over an 18-year period. The coding of news text started with first identifying the objects from an ontology of predefined knowledge objects, which identified all actors and issues mentioned within the context of Dutch politics. Then, texts were converted into "nuclear sentences," which captured relationships between those objects, enabling flexible re-analysis of data for different research questions. Diversity was assessed in terms of both parties (entities) and issues (topics). In terms of party diversity, the number of nuclear sentences referencing a political party or any of its politicians was analyzed. For issue diversity, the number of nuclear sentences with an issue belonging to one of the issue dimension categories contemplated in the paper was taken into consideration. This was the other paper that mentioned normative frameworks for measuring diversity. Here, on top of measuring open diversity as explained, which consists in measuring open diversity according to the deliberative normative framework, reflective diversity was also measured according to the liberal-aggregative normative framework. Reflective normative diversity was assessed through horizontal diversity, which consists of comparing the relative news attention for each party type with the relative power base of these parties in Parliament prior to the elections.

Finally, Hashimoto et al. presented a graph-based approach to measure topic diversity in tweets [37]. The method developed by these authors was based on clustering a tweet graph using the Data Polishing algorithm [85, 86], which automatically determines the number of subtopics in a topic. In the first step, tweet-word matrices were constructed, as shown in Figure 3.12, where m corresponds to the number of tweets and n to the number of words. W_{mn}=1 if the tweet m contains the word n, 0 otherwise.

$$W = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & & \ddots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{pmatrix}$$

Figure 3.12: tweet-word matrix [37].

Then, a tweet graph was generated. Nodes are tweets, and they are connected by edges if Jaccard Similarity between them > 0.3. Finally, a Clustering tweet graph based on the previous one was built. Here, two vertices are connected in the new graph if and only if the vertices seem to belong to the same cluster, according to the following expression, which means that u and v belong to the same cluster, as their neighbor's sets must overlap above a certain threshold when they do:

$$\frac{|N[u] \cap N[v]|}{|N[u] \cup N[v]|} > \theta_p$$

Data polishing applies this graph reconstruction iteratively until it does not change. MACE algorithm [58] is used to obtain the clusters of the resulting graph by performing maximal clique enumeration. The resulting clusters are interpreted as sub-topics, and the number of clusters is interpreted as the diversity of the topic.

Finally, an effort to assess viewpoint diversity was developed by Kiritoshi and Ma [45]. In this paper, the Difference in Opinion (DO) calculation quantifies the extent of subjective differences in descriptions between two news articles reporting the same event. This measure uses polarities (positive, negative, or neutral) associated with named entities mentioned in the articles. The formula goes as follows:

$$do(a,o) = relmut(a,o) \times sup(a,o)$$

where relmut(a, o) measures how many named entities are shared between the two articles in relation to the total unique entities they mention and sup(a, o) computes the difference in polarity scores for each shared entity between the two articles.

3.4 Summary

The work presented in this section encompasses three of the key pillars of the study of the dynamics of the news ecosystem, namely variations in coverage, publication patterns, and media diversity. Having this in mind, the insights given by this related work analysis are notable.

First, this investigation provided vital insights into each of the aforementioned pillars and was essential to understand how different fields of informatics can be leveraged to improve the effectiveness of the study of each of these pillars. For instance, Machine Learning (ML) and DL methods are pivotal in evaluating media diversity, while techniques derived from NLP can be valuable for aiding in the detection of various types of elements in news documents. The field of HCI

is also essential, as it plays a crucial role in determining the most effective visualization methods to observe coverage variations and the evolution of publication patterns. On top of the more comprehensive understanding this analysis provides, it also identifies a notable gap in the study of the Portuguese online news ecosystem. In fact, the most explored pillar was the publication patterns [26, 25]. In terms of diversity, only topic diversity has already been explored [25].

In conclusion, the examination of previous work in this chapter was crucial, as it provided a comprehensive overview of the theme. This knowledge is essential for developing the solution presented in the next chapter.

Chapter 4

Solution

The solution developed to carry out the exploration of the Portuguese online news landscape is detailed in this chapter. In Section 4.1, the process of creating a PostgreSQL[1] database to accommodate the data contained in the dataset of news headlines and summaries is described and, in Sections 4.2 and 4.3, the data are evaluated and characterized. Finally, in Section 4.4, the processes of topic modeling, topic labeling, and entity extraction are described.

4.1 Database Creation

The first step to put the proposed solution into practice consisted of creating the database to accommodate the data. In order to do this, the first step consisted of downloading a dump psql file containing the database of news headlines and summaries. Following, a PostgreSQL database was created and fed with the dump file, by restoring the dump using psql. The database schema can be seen below in Figure 4.1.

Having the database created, everything from here on was orchestrated by a Python script. After connecting to the previously created PostgreSQL database, a connection to a DuckDB[2] database was performed, and the data from the tables contained in the PostgreSQL database were loaded into DuckDB tables. The choice of DuckDB is motivated by its ability to seamlessly integrate with external databases such as PostgreSQL, allowing data to be queried or transferred without the need for intermediate export/import steps [70]. This tight integration enables efficient analytical workflows by combining DuckDB's in-process execution model with the relational capabilities of PostgreSQL, minimizing data movement and optimizing SQL operations.

With the database created, the next section delves into the characterization and evaluation of the available data.

[1] https://www.postgresql.org/
[2] https://duckdb.org/

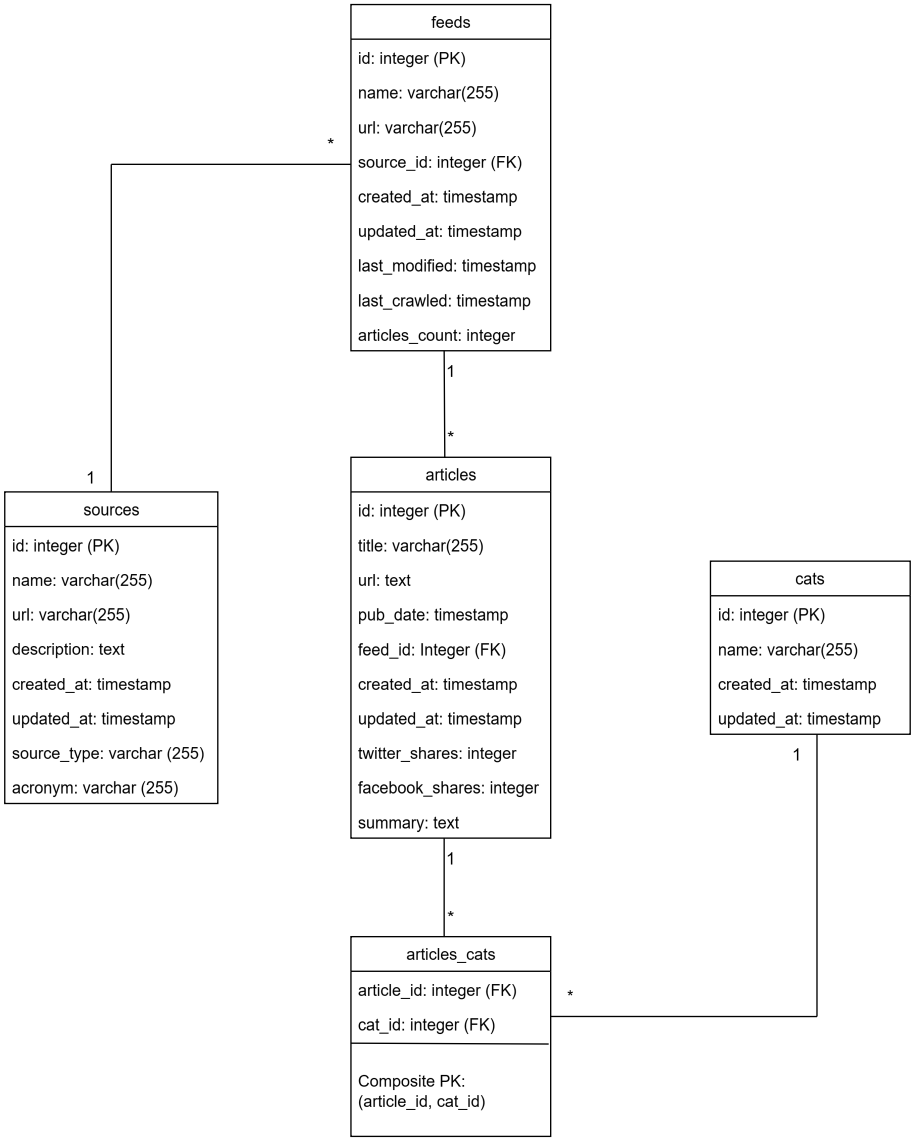

Figure 4.1: Database schema represented through a UML diagram.

4.2 Corpus Characterization

The provided dataset contains five tables: "articles", which is the main table, as it contains the news documents and several relevant information about them, such as the title, the summary, the publication date, and the feed to which they belong; "cats", which contains various topics or categories covered by the documents in the dataset; "articles_cats", which associates each article with one or more categories; "sources", which contains information about the sources in the dataset, such as their name, acronym and their type; and "feeds", which contains information about the feed from where each news article comes from. The UML diagram shown in Figure 4.1 supports this information. Listings 4.1, 4.2, 4.3, 4.4, 4.5 show an example entry from each of these tables.

```
[
    {
        "index": 1,
        "title": "Neves Adelino demite-se da administração da Cimpor",
        "url": "http://expresso.sapo.pt/neves-adelino-demite-se-da-administracao-da
            -cimpor=f901465",
        "pub_date": "2014-12-05 20:43:00",
        "feed_id": 3,
        "created_at": "2014-12-06 01:00:04.209931",
        "updated_at": "2016-09-09 19:38:06.298019",
        "twitter_shares": 6,
        "facebook_shares": 9,
        "summary": "Renúncia a 26 de novembro foi esta sexta-feira comunicada a
            CMVM."
    }
]
```

Listing 4.1: Example entry from the `articles` table.

```
[
    {
        "index": 2,
        "name": "mundo",
        "created_at": "2014-12-06 01:23:32.405942",
        "updated_at": "2014-12-06 01:23:32.405942"
    }
]
```

Listing 4.2: Example entry from the `cats` table.

```
[
    {
        "article_id": 52380,
        "cat_id": 8071
    }
]
```

Listing 4.3: Example entry from the `articles_cats` table.

```
[
    {
        "id": 2,
        "name": "Expresso",
        "url": "http://expresso.sapo.pt",
```

```
        "description": null,
        "created_at": "2014-10-22 09:56:03.842251",
        "updated_at": "2015-01-08 14:35:31.410967",
        "source_type": "national",
        "acronym": "expresso"
    }
]
```


Listing 4.4: Example entry from the `sources` table.


```
[
    {
        "id": 3,
        "name": "Economia",
        "url": "http://expresso.sapo.pt/static/rss/economia_23413.xml",
        "source_id": 2,
        "created_at": "2014-10-22 09:56:03.84711",
        "updated_at": "2024-10-25 15:53:55.298684",
        "last_modified": "2014-12-09 17:15:24",
        "last_crawled": "2024-10-25 15:53:55.29777",
        "articles_count": 22309
    }
]
```


Listing 4.5: Example entry from the `feeds` table.

The dataset consists of 13,236,588 news headlines and summaries published by 68 different media sources of five different types: national, which corresponds to news published by Portuguese media sources; international, which consists of news published by media sources outside of Portugal; research; archive; and blogs. The majority of news sources are national (27, or 39.71%). This study focuses on Portuguese news, and, actually, the majority of news comes from national media sources (40.87% or 5,410,268 articles). This information can be seen in Table 4.1. For simplicity and due to time constraints, only news published by Portuguese media sources in 2024 was considered for topic and entity extraction, which consists of a set of 223,975 articles (which corresponds to 32.63% of the total number of articles published in 2024) published by 14 sources from January 5th to October 25th. Considering the overall distribution of articles published in that year, three sources (21.43%) published less than 10,000 articles, six sources (42.86%) published between 10,000 and 20,000 articles, three sources (21.43%) published between 20,000 and 30,000 articles, and two sources (14.29%) published more than 30,000 articles. This distribution of the number of news published by source can be seen in Figure 4.2.

Table 4.1: Sources per type and news per source type, in absolute numbers and percentages

	Sources per Type	Sources per Type (%)	News per Source Type	News per Source Type (%)
national	27	39.71%	5,410,268	40.87%
international	14	20.59%	3,902,872	29.49%
research	13	19.12%	2,171,711	16.41%
archive	11	16.18%	1,694,216	12.8%
blogs	3	4.41%	57,521	0.43%

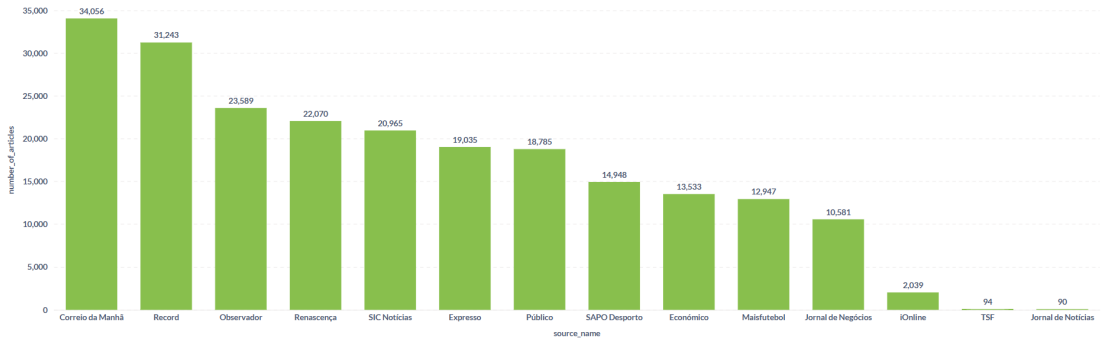

Figure 4.2: News per Portuguese source from January to October 2024.

Another perspective from which the collection can be characterized is related to the headlines' and summaries' sizes. Considering that, in Subsection 4.4.3, a model based on BERTimbau [81] is used for NER, and BERTimbau as a limit of 512 tokens it can process at a time, it is important to assess the number of headlines and summaries with more than 512 tokens. In the case of headlines, there are no headlines that exceed this limit. However, in the case of summaries, there are 238,487 news documents with summaries that contain more than 512 tokens. This information already gives a hint of the preprocessing needed to apply the BERTimbau-based model to the news summaries in the dataset.

4.3 Corpus Evaluation

The origin of the documents presented in this dataset of news headlines and summaries is either the RSS or Atom feeds of the sources. Thus, the dataset has some flaws that are explored here in this section.

First, data evaluation was performed taking into account three different measures: completeness, which checks for null values in each of the columns of each of the tables present in the dataset; uniqueness, which checks for duplicate records in each table; and validity, which searches for invalid values. The most important observations to highlight are that, in terms of completeness, there are 595,612 news articles (4.50%) that do not have any summary and, concerning uniqueness, there are two repeated records in the articles table. Regarding validity, the dates and timestamps were checked to ensure they were in the correct format and did not refer to future

dates. Additionally, other column values were validated to confirm they were less than zero. No issues were found.

On top of the conclusions drawn from this analysis, a notable inconsistency was found in the dataset: some news articles have publication dates earlier than their creation dates (12,174,759 or 91.98%), and update dates preceding their publication dates (870,469 or 6.58%). These inconsistencies stem from the fact that the articles are sourced from the RSS or Atom feeds of the media outlets. These feeds typically prioritize providing content quickly over maintaining strict consistency in metadata fields, such as the creation and update timestamps, leading to irregularities in the recorded publication chronology [49].

Furthermore, an additional observation concerned the textual cleanliness of headlines and summaries. Taking the following summary as an example: "(Reuters) — Spanish rider Alex Rins has signed a two-year extension with Suzuki, the MotoGP team announced on Thursday ahead of the French Grand Prix at Le Mans.", it can be seen that it contains text that clearly does not belong to the summary of the news document. (Reuters) corresponds to the name of the source this news came from and, consequently, does not belong to the summary itself. So, there needs to be a concern with cleaning the text, mainly taking into consideration that entities and topics are going to be extracted from these news documents. If agency attributions such as this one are maintained, entity extraction tools would recognize "Reuters" as an Organization, and the entity extraction results would be compromised, as this entity does not belong to the news article's summary.

In the following sections, the processes for extracting topics and entities from the documents are described.

4.4 Topic and Entity Extraction

In order to extract topics and entities from the news headlines and summaries contained in the dataset, it is necessary to develop an adequate method. Taking this into account, the processes of topic modeling and topic labeling are first described in Subsections 4.4.1 and 4.4.2. Then, entity extraction is also explained in Subsection 4.4.3.

As has already been explained in Section 4.3, it is necessary to clean the data, and this step is common to both topic and entity extraction. Prefixes and suffixes commonly found in news articles, such as agency attributions, were removed to retain only the core content. An example of this can be seen in the following summary: "(Reuters) — Spanish rider Alex Rins has signed a two-year extension with Suzuki, the MotoGP team announced on Thursday ahead of the French Grand Prix at Le Mans." Specific recurring phrases that added no informational value were also eliminated, such as "Continue reading", for example. HTML entities were decoded using the html Python library[3], and various typographical inconsistencies were corrected, including normalizing dashes, fixing curly quotes, and replacing bullet points. On top of these, additional spaces at the beginning and end of sentences were eliminated, and punctuation inconsistencies were addressed,

[3] https://pypi.org/project/html/

ensuring proper spacing and readability, with the aid of regular expressions. Unicode normalization was applied to resolve encoding issues using the unicodedata Python library[4]. These steps collectively refined the text, making it cleaner and more structured for further processing.

4.4.1 Topic Modeling

Topic modeling is used in information retrieval to infer the hidden themes in a collection of documents and thus provides an automatic means to organize, understand, and summarize large collections of textual information [1]. The main aim here is to group the different documents into clusters according to the topic they are covering. With this objective in mind, three topic modeling tools were chosen to be compared: LDA [10], BERTopic [32], and GSDMM [96]. The motivation behind these choices lied in selecting a representative model from each category: in the case of LDA — the traditional model for topic modeling — a probabilistic model, in the case of BERTopic, a neural model based on the Bidirectional Encoder Representations from Transformers (BERT) architecture, and, finally, in the case of GSDMM, a model based on Dirichlet Multinomial Mixture (DMM) [40], which is also tailored for short texts, in opposition to the other two models, which are tailored for generic texts, although being used in this context with short texts. It is also of interest to assess whether the model specialized in short texts performs better than the other models, as, in this case, we are dealing with short texts — news headlines and summaries.

Although topic extraction was performed in all news published by national news sources in 2024 in 5.2, in order to evaluate which of these models has better performance, the subset of news published on just a day of October 2024 was used, specifically October 25th, 2024. This day was chosen because it is the most recent day present in the dataset. Since the news published on October 25th is to be used in the user survey to evaluate the topic modeling tools and the topic labeling strategies in Section 4.4.2.1, it is better if the news is the most recent possible, so users can better recall the events that occurred, which may help provide more reliable feedback. With the data selected and cleaned, the next step consists of applying the topic modeling tools.

4.4.1.1 BERTopic

BERTopic was implemented using the bertopic Python library[5], and the news documents were processed per day in 2024. First, the title and summary of each news article were concatenated into a single string, with a period used as a separator between them, so the model could be fed with more information about each news document. BERTopic combines sentence encoders, implemented using the sentence_transformers Python library[6], to embed documents into dense vectors, — in this case, the one used was PORTULAN/serafim-100m-portuguese-pt-sentence-encoder-ir[7] — with UMAP, implemented using the umap Python library[8], which reduces high-dimensional

[4] https://docs.python.org/3/library/unicodedata.html
[5] https://pypi.org/project/bertopic/
[6] https://pypi.org/project/sentence-transformers/
[7] https://huggingface.co/PORTULAN/serafim-100m-portuguese-pt-sentence-encoder-ir
[8] https://pypi.org/project/umap-learn/

embeddings to 2D for clustering, and HDBSCAN, implemented using the hdbscan Python library[9], which clusters the reduced vectors into topic groups. Also, a vectorizer model was used to create topic representations without stop words. The stop words were obtained from the nltk Python library[10] and unigrams were used to create the topic representations. Bigrams and trigrams were also tested, but unigrams resulted in a better coherence CV score [74]. This metric evaluates how similar the context of each word that represents a topic is to the others in that topic, under the assumption that words in a coherent topic will appear in similar contexts in a large corpus. Thus, a coherent topic is one where the top words strongly relate to each other in meaning. The coherence CV was also the metric used to fine-tune all the hyperparameters of the UMAP and HDBSCAN, and it was implemented using the CoherenceModel module from the gensim Python library[11].

With these decisions made, the model was trained using documents composed of each article's title and summary — and returned the topics for each document and the probability of the topic being assigned to that document. Weakly assigned documents — documents with topics assigned with a probability less than 0.3 — were considered outliers. All values from 0 to 0.5 were tested, with increments of 0.1, but 0.3 was the one that obtained a better coherence CV score. BERTopic assigns a number starting from 0 to each news document which works as the topic id. In the case of outliers, a -1 was attributed to them to identify them as outliers. This way, the clusters were created, with documents with equal topic IDs being assigned to the same cluster. Finally, BERTopic assigns an arbitrary number of words to represent each topic. In this case, five was the chosen number. The model was executed multiple times on the October 25th dataset to examine the number of topic clusters formed and the assignment of news articles to those clusters. Every execution consistently produced between 20 and 30 topics. This pattern influenced the decision to use a similar range of topic numbers in experiments conducted with the other topic modeling tools, which require a number of topics to be specified.

4.4.1.2 LDA

LDA is implemented using the gensim library, and news documents were also processed per day. The title and summary of each news article were also merged into one string. The preprocessing pipeline for LDA included tokenization, lemmatization, stop words, and punctuation removal, as the implementation used here was based on the implementation provided by Babaloa et al. [7]. The stop words were obtained through the nltk Python library. Next, bigrams and trigrams were created from each of the tokenized strings of headlines and summaries to help capture multi-word expressions, and a Gensim dictionary of words was created, where too rare — words that appear in fewer than two documents — and too common words — words that appear in more than 50% of the documents — were filtered out. A special corpus was created by converting each string into a bag-of-words representation using the created dictionary. Then, the LDA model was trained on the corpus with 20 topics, 20 passes, which are the times the model goes through the corpus, and

[9] https://pypi.org/project/hdbscan/
[10] https://pypi.org/project/nltk/
[11] https://pypi.org/project/gensim/

200 iterations, which represents the number of iterations per pass — all these hyperparameters were also fine-tuned with the help of the Coherence CV measure. For the number of topics, values from 20 to 30 with increments of five were tested, with 20 getting the best coherence CV score. For the number of passes, values from 10 to 30 were tried out, and for the number of iterations, values from 50 to 500 with increments of 50 were used. In the case of the number of passes, the coherence CV score was found to stabilize between 20 and 30. Therefore, a value of 20 was selected as it provided comparable performance while reducing computational time. Similarly, for the number of iterations, increasing beyond 200 did not lead to a significant improvement in coherence, while resulting in longer training times. Thus, 200 iterations were chosen as a balanced trade-off between model quality and computational efficiency. Finally, for each of the 20 topics, the top five words that best represented them—according to the LDA model—were extracted. Similar to BERTopic, LDA assigns a score to each word associated with a topic, where a higher score indicates a stronger representation of the topic. The LDA model was then used to determine the most probable topic for each news document, based on the learned topic distribution. In this way, news documents assigned to the same topic were grouped together, forming the final clusters.

4.4.1.3 GSDMM

GSDMM is implemented using the gsdmm library[12], and its implementation process was similar to that of LDA. The only difference is that, in the case of GSDMM, the tokenized strings of headlines and summaries from which bigrams and trigrams were created only contained nouns, proper nouns, verbs, adjectives, and adverbs, as a result of the POS-based lemmatization process that was also applied. In addition to this, GSDMM was trained with 25 instead of 20 topics — similar to what happened in the case of LDA, two other experiments were also conducted with 20 and 30 topics, but 25 was the number of topics that resulted in the best coherence CV score — an alpha and beta values of 0.1 [96] and with 10 iterations, as the GSDMM model is fast to converge [96]. As happened with the other two models, the top five words that represented the topics and that are given by GSDMM were also extracted and the model was used to obtain the most probable topic for each news document based on the distribution of topics learned.

4.4.2 Topic Labeling

Topic labeling is the process of assigning descriptive labels to topics that emerge from a topic modeling algorithm [28]. This is an important step in the topic extraction process because the representations produced by topic modeling tools are not always easily interpretable by humans [28]. Therefore, it is crucial to find a way to represent these topics in a way that is easily understandable. Taking this into consideration, two different topic labeling approaches are proposed. The first strategy consists of using the top five words that represent each cluster and that can be obtained from the tool where the cluster comes from. The second strategy involves using Yake![13].

[12] https://github.com/rwalk/gsdmm
[13] http://yake.inesctec.pt/

Yake! is a light-weight unsupervised automatic keyword extractor that uses statistical text features extracted from single documents to select the most relevant keywords of a text, and is capable of extracting keywords from texts independently of language and domain [16]. For this effect, the yake library for Python was used[14], with the language set to Portuguese, max_ngram_size set to 1, as the objective was to extract unigrams, and num_keywords_per_headline set to 5, as the purpose was to extract the top five keywords that best represent each cluster. Each cluster was given as a list to Yake!, which returned, for each, the top five keywords that represent it.

Following, the user survey used to evaluate each of the three topic modeling tools and each of the two topic labeling strategies is detailed.

4.4.2.1 Topic Modeling and Topic Labeling User Survey

Before delving into the user survey, it is worth mentioning the manual evaluation that was conducted prior to its development. As already explained in Subsection 4.4.1, all the news documents published on October 25th were grouped into clusters by each of the topic modeling tools. To have a baseline for comparison with the user survey results, a manual evaluation was carried out. For each cluster, the news documents were reviewed individually, and it was decided whether each one belonged to the cluster or not by comparing its headline with the top five words representing the cluster. This way, the hit rate for each topic modeling tool could be calculated by the following formula:

$$Hitrate = \frac{correctly\_assigned\_news}{total\_number\_of\_news} \times 100$$

Table 4.2 shows the hit rate for each topic modeling tool.

Table 4.2: Hit rate for each of the topic modeling tools.

BERTopic	LDA	GSDMM
64.39%	54.47%	58.50%

On top of this manual evaluation, a user survey was also created, with the objective of gathering insights from more people. The user survey was implemented as an online questionnaire using Google Forms[15]. The purpose of this form was to evaluate the performance of the three topic modeling tools. For the evaluation to be fair, the chosen news documents had to be the same for the three topic modeling tools. For this effect, the news headlines from three random clusters formed by BERTopic were selected for inclusion in the user survey, so the news documents were the same for the three tools. In the case of the topic modeling process performed by GSDMM, these news documents also belonged to three different clusters, and, in the case of LDA, these same news documents belonged to four different clusters. Taking this into account, the user survey ended up with 10 news clusters: three for BERTopic, three for GSDMM, and four for LDA. All these clusters were composed of three or more news. For each news headline in each cluster, the

[14] https://pypi.org/project/yake/
[15] https://docs.google.com/

user was asked to evaluate how well it fit the overall theme of the set, on a scale from "does not fit at all" to "definitely fits". Figure 4.3 shows an example of one of these questions.

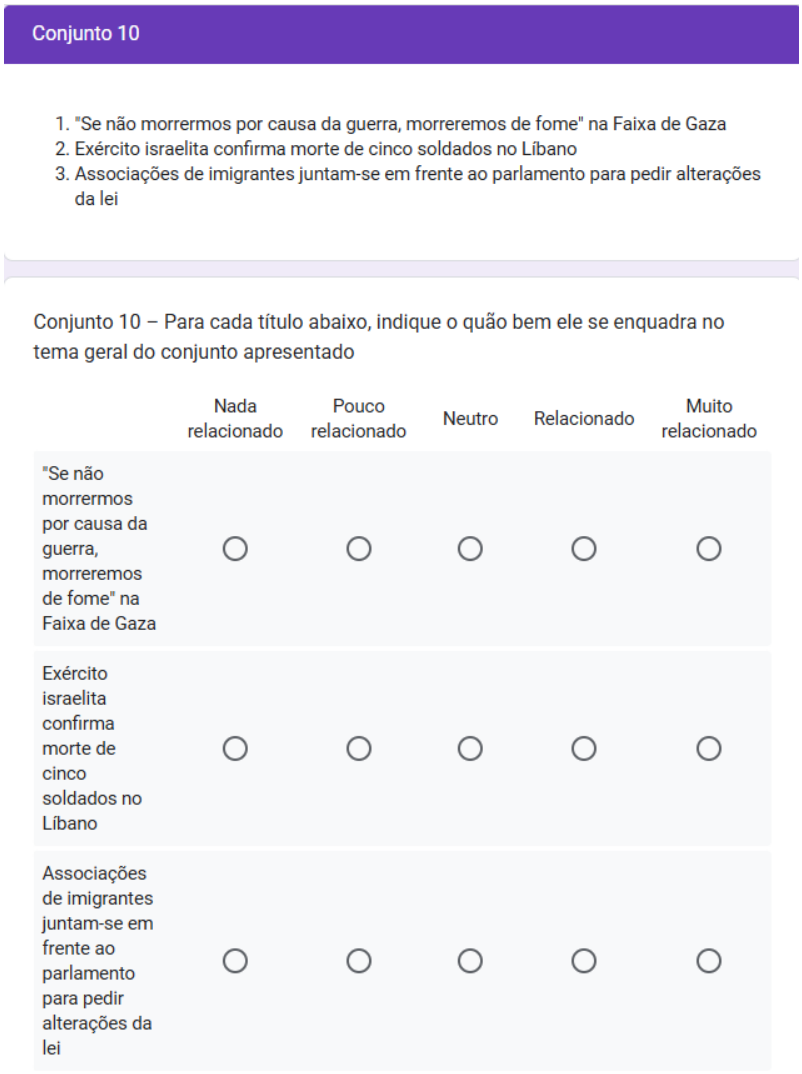

Figure 4.3: Question to evaluate the topic modeling tools.

Besides evaluating the three topic modeling tools, the user survey also assessed the two topic labeling strategies previously described in Subsection 4.4.2. In the user survey, two sets of labels for each cluster were presented for the user to evaluate: the set of five words extracted by the topic modeling tool where the respective cluster comes from and the set of five keywords given by Yake! for that cluster. The user was asked to evaluate how well each of the sets of keywords represented the content of the cluster, according to the news included in it, on a scale from "not representative at all" to "highly representative". Figure 4.4 shows an example of one of these questions.

After explaining the process of selecting the better method for topic extraction from the news documents in the dataset, the next step is to outline the procedure for identifying the most effective tool for entity extraction. The following section explores this process in detail.

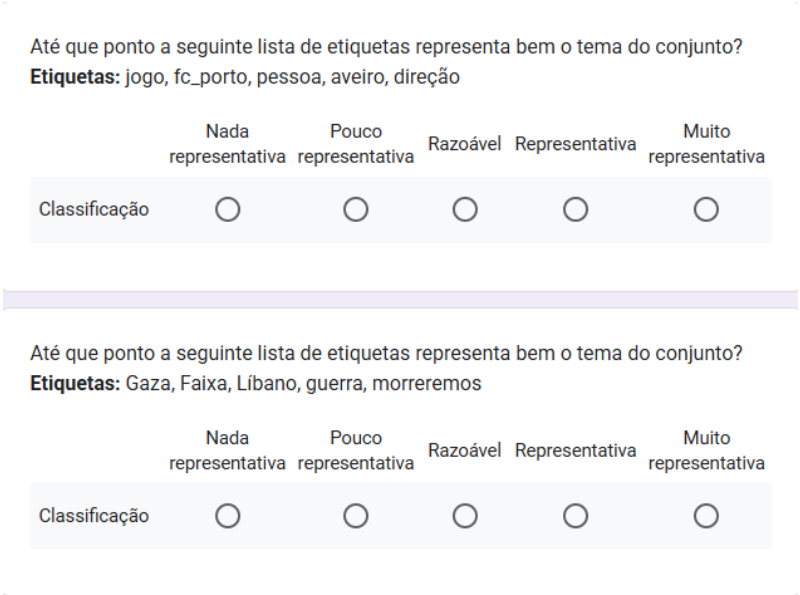

Figure 4.4: Question to evaluate the topic labeling strategies.

4.4.3 Named Entity Recognition

In the domain of NLP, NER stands out as a pivotal mechanism for extracting structured insights from unstructured text and plays a crucial role in information extraction and knowledge organization [68]. It involves identifying and classifying named entities — such as individuals, organizations, locations, dates, and more — within textual data. Its key role lies in extracting structured information from unstructured text, thereby improving data retrieval, analysis, and comprehension [68].

So, considering the objective of extracting Person, Organization, Location, Event, Product, and Date entities, the main concern here consists in choosing the most adequate NER tool to extract these types of entities. For this effect, several experiments were done to understand which tool performed better in entity extraction from news headlines and summaries.

4.4.3.1 Entity Extraction Tools Evaluation

The first step in evaluating the quality of different NER models consisted of selecting which ones to test. The ones selected were spaCy[16], GLiNER [98], and a model fine-tuned for NER based on BERTimbau [81]. The choice of spaCy was motivated by its availability of models capable of detecting entities in multiple languages, including both a Portuguese-specific model and a multilingual model, both of which were tested. The spaCy models used were:

- Portuguese models[17]:
  - pt_core_news_sm

[16] https://spacy.io/
[17] https://spacy.io/models/pt

  - pt_core_news_md
  - pt_core_news_lg

- Multi-language models[18]:
  - xx_ent_wiki_sm

The choice of the neural model GliNER was motivated by the fact that the entities' labels are user-chosen, and because it is a model that performs well in zero-shot context [98]. Also, it includes a multi-language model. The GLiNER model used was:

urchade/gliner_multi[19]

Finally, the model based on BERTimbau was chosen because it was trained on the HAREM dataset[20], and, thus, is a neural model specifically tailored for Portuguese. The model chosen was:

marquesafonso/bertimbau-large-ner-total[21]

After choosing the models, the next step consisted of applying them to a subset of news headlines and summaries. Although in 5.2 entity extraction was performed in all news published by national news sources in 2024, for evaluation purposes only 25 news were randomly selected here, one from each day of October 2024 when there was news published (from October 1st to October 25th), so a manual evaluation could be performed. For each of the 25 news documents, entities were extracted from both the news title and summary.

In the case of spaCy, the models only have the labels "PER" (Person), "LOC" (Location), "ORG" (Organization), and "MISC" (Miscellaneous). The "MISC" label was excluded from consideration due to its vague and undefined scope. The situation differs for both GLiNER and the BERTimbau-based model. In GLiNER, entity labels are user-defined. For the purposes of the manual evaluation — aligned with the objective outlined in Subsection 4.4.3 — the following labels were selected: "Person", "Location", "Organization", "Date", "Product", and "Event". In contrast, the BERTimbau-based model uses a fixed set of 10 entity labels: "PERSON", "LOCATION", "ORGANIZATION", "EVENT", "TIME", "WORK OF ART", "VALUE", "ABSTRACTION", "THING", and "OTHER". Among these, the labels selected for evaluation were: "PERSON", "LOCATION", "ORGANIZATION", "EVENT", "WORK OF ART", and "TIME". While "PERSON", "LOCATION", and "ORGANIZATION" directly correspond to the equivalent GLiNER labels, "WORK OF ART" and "TIME" do not. However, they were identified as the closest approximations to "Product" and "Date" in GLiNER, respectively. Although it would be possible to modify the GLiNER configuration to match BERTimbau's labels, doing so would not align with the evaluation objective defined in Subsection 4.4.3. The correspondence between the entity labels used in spaCy, GLiNER, and the BERTimbau model is summarized in Table 4.3.

[18] https://spacy.io/models/xx
[19] https://huggingface.co/urchade/gliner_multi
[20] https://huggingface.co/datasets/Linguateca/harem
[21] https://huggingface.co/marquesafonso/bertimbau-large-ner-total

Table 4.3: Entities correspondence between spaCy, GLiNER, and the BERTimbau-based model.

spaCy	GLiNER	BERTimbau-based
PER	Person	PERSON
ORG	Organization	ORGANIZATION
LOC	Location	LOCATION
	Event	EVENT
	Date	TIME
	Product	WORK OF ART

An important aspect to note is that the BERTimbau-based model has a strict limit of 512 tokens that it can process at a time [81]. And, as explained in Section 4.2, there are 238,487 news documents with summaries that contain more than 512 tokens. Thus, one concern that had to be taken into account was that any summary that contained more than 512 tokens had to be split into chunks of that size and each chunk had to be processed individually.

After applying the models to each of the news headlines and summaries, the results were stored in a JSON file. An example entry from this file can be seen in Listing 4.6.

```
[
    {
        "index": 1,
        "title": "S4Congress - Desafios emergentes e a Segurança dos Eventos
            Desportivos",
        "summary": null,
        "entities": {
            "sm_pt" : [
                [
                    "Desafios",
                    "LOC"
                ]
            ],
            "md_pt": [],
            "lg_pt": [],
            "xxsm": [
                [
                    "S4Congress - Desafios",
                    "PER"
                ]
            ],
            "gliner_xx": [
                [
                    "S4Congress",
                    "Event"
                ]
            ]
            "BERTimbau": [
                [
```

```
                    "S4Congress",
                    "ACONTECIMENTO"
                ],
                [
                    "dos Eventos Desportivos",
                    "ACONTECIMENTO"
                ]
            ]
        }
    }
]
```

Listing 4.6: Example of a news document from the JSON file with entities extracted by all the tools.

To determine the most suitable tool, the F1-score metric was employed, calculated using the following formula:

$$F1-score = \frac{2 \times TP}{2 \times TP + FP + FN} \times 100$$

where "TP" are the True Positives, which are words correctly identified as entities and correctly labeled, "FP" are the False Positives, which are words wrongly identified as entities or wrongly labeled, and "FN" are the False Negatives, which are words that are entities, but are not identified as so. This formula allowed to conclude which tool was the most effective in what concerns the extraction of entities.

Based on this, two evaluations were conducted. The first involved comparing the spaCy models with each other, and the two neural models with each other. The objective of this first evaluation was to compare the two sets of similar named entity extraction tools, which extract the same types of entities. In order to perform these two evaluations, the correct entities were first manually identified and labeled according to NER annotation rules [35, 84]. The explanation of these rules can be found in Table 4.4.

Table 4.4: Rules for identification and labeling of the different entity types [35, 84].

Entity Type	**Description**
Person	- Real or fictional people - A named group of people, e.g., music bands (e.g., The Beatles) - Religious figures (e.g., God)
Organization	- Provider of products or services - Includes companies, hospitals, schools, universities, political parties, unions, police, gendarmerie, churches, armed forces (named), sports clubs, etc.
Location	- Districts, towns, villages, cities - Regions, municipalities, departments, provinces, etc. - Countries - Continents - Mountains, plains, plateaus, caves, volcanoes, canyons - Oceans, seas, rivers, streams, lakes, swamps - Planets, stars, galaxies - Roads, highways, streets, avenues, squares, etc. - Buildings - Physical addresses - Electronic contact information (phone numbers, fax numbers, URLs, email addresses, etc.)
Product	- Newspapers, magazines, broadcasts, sales catalogs, movies, etc.
Date	- An absolute date (specific date, not a relative one; dates that include only a day and month, a month and year, or just a year or century) - A time span between two absolute dates
Event	- A concrete event, specific in time and space

After annotating the entities, the number of "TP", "FP", and "FN" were counted, and the F1-score formula was applied to each of the tools. Obviously, the spaCy models were only evaluated in relation to the entity types Person, Location, and Organization, as these are the only entity types they can identify. On the other hand, the GLiNER model and the BERTimbau-based model were evaluated regarding all types of entities. The evaluation of the news document displayed in Listing 4.6 can be seen in Table 4.5 for exemplification purposes.

Table 4.5: Evaluation of one news document with number of "TP", "FP", and "FN".

Entities: S4Congress – Event						
Metric	sm_pt	md_pt	lg_pt	xxsm	gliner_xx	BERTimbau
TP	0	0	0	0	1	1
FP	1	0	0	1	0	1
FN	0	0	0	0	0	0

As shown in the table, the spaCy models do not receive an "FN" for failing to identify "S4Congress" as an entity, since they are not designed to recognize this type of entity. However, they are assigned an "FP" if they incorrectly identify a non-entity as an entity, as is the case with the "sm_pt" and "xxsm" models in Figure 4.6, or if they mislabel an entity. Finally, the second evaluation consisted in comparing the better spaCy model with the better neural model. This process is explained in further detail in the next chapter.

4.5 Summary

This chapter describes the practical implementation developed to explore the Portuguese online news dataset with regard to publication patterns, topic, and entity extraction.

To support this, the following methodology was followed. First, a PostgreSQL database was created and loaded with preexisting news headlines and summaries. For processing and querying efficiency, the data were mirrored in a DuckDB instance that supports fast SQL analytics. The dataset was then characterized and evaluated, and five tables were used: articles, cats, articles_cats, sources, and feeds. These were examined in terms of structure, content, and relevance to the research goals. For topic extraction, three models were evaluated: BERTopic, LDA, and GSDMM representing neural, probabilistic, and DMM-based approaches respectively. Two approaches were tested for topic labeling: one using the word representations provided by each of the three topic modeling tools, and the other employing a keyword extractor — Yake! — to generate topic labels. A manual evaluation was conducted to compare the three topic modeling tools with each other and a user survey was also conducted to compare the same topic modeling tools and the two topic labeling strategies. For entity extraction, multiple NER tools were evaluated: spaCy, GLiNER, and a BERTimbau-based model. A manual evaluation using a metric called F1-score guided the final choice of models.

The solution outlined in this chapter provides the basis for the empirical analysis conducted in Chapter 5.

Chapter 5

Results and Discussion

The results of the exploration of the data contained in the dataset of news headlines and summaries are presented in this chapter alongside the analysis of these same results. In Section 5.1, the results of the user survey used to evaluate the different topic modeling tools and topic labeling strategies are presented alongside the results of the manual evaluation performed to decide on the most effective tool for entity extraction, and, in Section 5.2 the enrichment of the dataset with the topics and entities extracted is detailed. Section 5.3 provides an overview of Metabase, a data visualization tool used here to facilitate a deeper understanding of the dataset. The comparison of publication patterns for two different sources, according to distinct temporal granularities, and the evolution of coverage across different topics, entities, and entity types is also explored in Sections 5.4 and 5.5, respectively.

5.1 User survey and Entity Extraction Tools Evaluation Results

As mentioned in Chapter 4, two evaluations were conducted to assess which was the most adequate entity extraction tool to use. The results of the first evaluation can be seen in Table 5.1. As can be observed in the table, the multilingual spaCy model obtained the better performance (77.89%), while the BERTimbau-based model surpassed the GLiNER multilingual and achieved a better F1-score — 71.43% for the former against 64.41% for the latter.

Table 5.1: F1-score for each of the entity extraction tools (first evaluation).

sm_pt	md_pt	lg_pt	xxsm	gliner_xx	BERTimbau
57.73%	68.13%	71.58%	77.89%	64.41%	71.43%

The second evaluation was conducted between the spaCy multilingual model and the BERTimbau-based model — the former being the one which obtained the better score of all the spaCy models, and the latter the one which obtained the better score of the two neural models. This evaluation was made considering only Person, Location, and Organization entity types. The objective was to understand whether spaCy multilingual achieved a better F1-score than the BERTimbau-based model for these types of entities. The results of the second evaluation can be seen in Table 5.2. As

can be observed, the multilingual spaCy model achieved a better score for these entity types than the BERTimbau-based model. This way, a combination of these two models was used: spaCy multilingual for Person, Location, and Organization, and the BERTimbau-based model for EVENT, WORK OF ART, and TIME. In case the BERTimbau model achieved a better results than spaCy for Person, Location, and Organization entities, only this model would be used for all the six types of entities.

Table 5.2: F1-score for each of the entity extraction tools (second evaluation).

xxsm	BERTimbau
77.89%	74.47%

Talking now about the results of the user survey on topic modeling and labeling, the user survey received 20 responses from 20 participants. All participants were university students. The conclusions drawn from this user survey are considerably different from the observations obtained through the manual evaluation. Focusing on the topic modeling part of the user survey, Figure 5.1 shows the results for one of the four LDA clusters used in the user survey. The blue bar corresponds to a rating of 1, red to 2, yellow to 3, green to 4, and purple to 5.

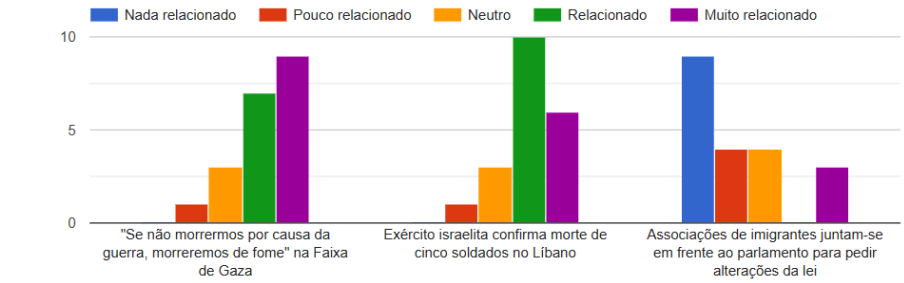

Figure 5.1: User survey results for one of the LDA clusters.

In order to assess what was the most adequate tool, a weighted mean needed to be calculated. The general formula for this calculation goes as follows:

$$S = \frac{1}{C \times N} \sum_{i=1}^{C} \left(\sum_{j=1}^{5} R_{i,j} \cdot j \right)$$

Explaining the formula, S represents the final score for the cluster, C represents the number of news documents in the cluster, and N corresponds to the number of participants in the user survey. R_i, j is the number of times document i received a rating of j from participants and is the possible rating value, ranging from 1 to 5.

Thus, considering the data shown in Figure 5.1, the calculation goes as follows:

$$\frac{(1\times 2+3\times 3+7\times 4+9\times 5)+(1\times 2+3\times 3+10\times 4+6\times 5)+(9\times 1+4\times 2+4\times 3\times 5)}{20\times 3}$$

Explaining the calculation, the first news document received one rating of 2, three of 3, seven of 4, and nine of 5. The second received one of 2, three of 3, ten of 4, and six of 5. Finally, the third received nine of 1, four of 2, four of 3, and three of 5. Also, as 20 participants responded to the user survey and as this cluster contains three news documents, this weighted mean had to be divided by 20 times 3, in order to normalize the calculation independently of the number of news in each cluster. Finally, as there is one score per cluster, the final score was calculated by adding all scores for each cluster and dividing this sum by the number of clusters for each tool. Using this weighted average, the most adequate topic modeling tool was chosen. The results are shown in Table 5.3, which allows to conclude that GSDMM is the tool that obtained the best score.

Table 5.3: Scores for each topic modeling tool obtained by the calculation of the weighted mean of the evaluations given by the participants.

BERTopic	LDA	GSDMM
4.09	3.61	4.49

The differences observed between the manual evaluation and the user survey results for topic modeling may be attributed to several factors. As the manual evaluation was conducted by the author, who is more familiar with the dataset and underlying themes, there may have been a deeper contextual understanding influencing the assessment. In contrast, the user survey was limited to a subset of clusters to ensure it remained concise and accessible to participants. It is possible that evaluating all clusters could have led to different outcomes, but due to the nature of user surveys, the scope had to be constrained.

Now focusing on the topic labeling part of the user survey, Figure 5.2 shows the results for the Yake! labeling strategy for one of the three GSDMM clusters used in the user survey. The blue bar corresponds to a rating of 1, red to 2, yellow to 3, green to 4, and purple to 5.

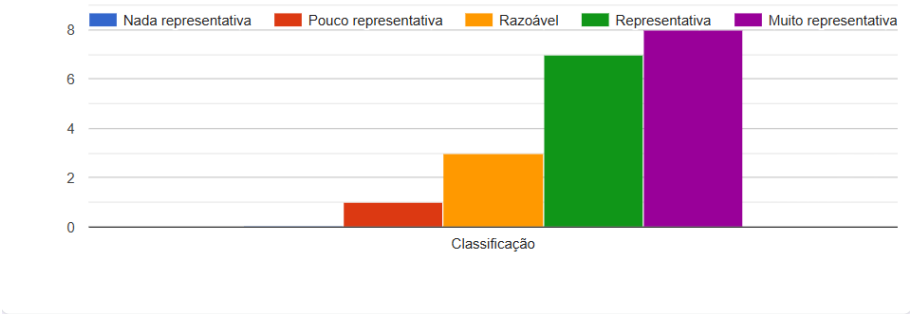

Figure 5.2: User survey results for one of the GSDMM clusters.

In order to assess what was the better labeling strategy, a weighted mean also had to be calculated. It is worth noting that these calculations were only made for GSDMM because it would not make sense to let the scores of the two topic labeling strategies for the other two topic modeling tools influence the scores for GSDMM specifically. So, considering the data shown in Figure 5.2,

the calculation goes as follows:

$$\frac{1\times 2+3\times 3+7\times 4+9\times 5}{20}$$

Explaining the formula, the Yake! strategy received one rating of 2, three of 3, seven of 4, and nine of 5. This mean was also divided by the 20 participants who responded to the user survey. Finally, as there is one score per cluster for each labeling strategy, the final score for each labeling strategy was calculated by adding all scores for each cluster and dividing this sum by the number of clusters for each tool. Using this weighted average, the most adequate topic modeling tool was chosen. The results are shown in Table 5.4. As can be observed, the Yake! strategy is the one with the best results.

Table 5.4: Scores for each topic labeling strategy obtained by the calculation of the weighted mean of the evaluations given by the participants.

Yake!	GSDMM-given top words
3.57	3.17

It is important to clarify that, although Yake! is traditionally designed for keyword extraction at the individual document level, in this study it was applied at the cluster level by aggregating all news items within a topic cluster into a single input. This adaptation allowed Yake! to extract representative keywords that characterize the entire cluster, making it more comparable to the representative terms generated by topic modeling tools such as GSDMM. However, GSDMM's underperformance in topic labeling compared to Yake! may be attributed to the fact that it does not generate explicit labels, instead relying on the most frequent terms in each cluster, which are often generic or uninformative in short texts [96]. In contrast, Yake! uses statistical and positional features to extract more distinctive and meaningful keywords [16]. This may have led to clearer and more human-interpretable labels, which likely explains its superior performance in the user survey.

Having now decided on the most suitable set of tools to extract topics and entities from the news documents in the dataset, the next section delves into the enrichment of the dataset with the topics and entities extracted for all news documents published by national news sources in 2024.

5.2 Dataset Enrichment with Extracted Topics and Entities

After choosing the most adequate topic modeling tool, topic labeling strategy, and set of entity extraction tools to use, the next step is to delve into the process of enriching the dataset of news headlines and summaries with the topics and entities extracted from each of the news documents published by Portuguese news sources in 2024. In the case of topics, the extraction was performed almost in the same way as explained in Subsection 4.4.1. In order to add these topics to the PostgreSQL database, a two-step process was carried out. It is relevant to note that the dataset was

enriched with the topics extracted by BERTopic and by GSDMM, as the aim here is also to compare the two, as the first obtained better results in the manual evaluation and the second obtained better results in the user survey. First, for the case of BERTopic, the topics extracted by BERTopic were linked with the id of the news document in the dataset from which they were extracted. Next, a new column was created in the "articles" table in the DuckDB database for the topics. The "topics_BERTopic" column is of type "text" and contains the topic associated with each news document, represented by the top five words given by BERTopic, which are separated by commas. However, an important difference between the extraction process for October 25th performed previously in Chapter 4 and the extraction process here needs to be noted, as the implementation of BERTopic used here skipped days with fewer than 30 articles (which corresponded to 41 days, or 13.90% of the days from January 5th to October 25th). This had to be done because BERTopic could not process days when fewer than 30 news documents were published. There was no need to do this on October 25th because on this day more than 30 articles were published.

The same thing was done for the case of GSDMM, The "topics_GSDMM" column of type "text" was also created in the "articles" table in the DuckDB database, and contains the topic associated with each news document as extracted by GSDMM, represented by the top five words extracted by Yake!

Regarding entities, as already explained in Subsection 4.4.3, all entities of types Person, Organization, and Location were extracted using the spaCy multilingual model, while all entities of types Event, Product, and Time are extracted using the BERTimbau-based model. These documents were processed day by day in chunks of 100. Similar to what is shown in Subsection 4.4.3, the results of entity extraction here were also stored in a JSON file, as can be seen in Listing 5.1.

```
[
    {
        "index": 12606311,
        "title": "Natação: Angélica André medalha de bronze nos 10 km em águas
            abertas",
        "summary": "Ao alcançar o 3.o lugar na prova, Angélica André garantiu uma
            vaga para os Jogos Olímpicos de Paris.",
        "entities": {
            "xxsm": [
                [
                    "Angélica André",
                    "PER"
                ],
                [
                    "Angélica André",
                    "PER"
                ]
            ],
            "BERTimbau": [
                [
                    "Jogos Olímpicos de Paris",
```

```
                    "ACONTECIMENTO"
                ]
            ]
        }
    }
]
```

Listing 5.1: Example of a news document from the JSON file with entities extracted by the final tools.

In order to add these entities and their respective types to the PostgreSQL database, a two-step process was carried out too, similar to what happened in the case of topics. First, the entities and their associated types were linked with the id of the news document in the dataset from which they were extracted. Next, two new columns were created also in the "articles" table in the DuckDB database, one for the entities and one for the corresponding entity types. Both the "entities" and "types_of_entities" columns are of type "text" and contain, respectively, all entities and their corresponding types found in each news document, separated by commas. The order of the entities matches the order of the respective entity types.

Finally, as a common step to the enrichment with topics and entities, the data contained in the DuckDB database had to be transferred to the PostgreSQL database in order to be explored and analyzed. First, a new table called "articles_everything" was created in the PostgreSQL database, and it accommodates all the documents published by Portuguese news sources in 2024 contained in the "articles" table in the DuckDB database. It is important to note that, in addition to enriching the "articles" table in the DuckDB database with topics, entities, and entity types, two other columns were added here: "source_name" and "type_of_source". The first contains the source that published the news document and was derived by joining the "feeds" and "sources" tables, and the second contains the source's type — which is always "national", as only Portuguese news is being considered — and was obtained directly from the "sources" table. After all the necessary columns were added to the "articles" table, the data transfer was ready to be done, and it was performed in chunks of 1,000 documents each. An example entry from this "articles_everything" table can be seen in Listing 5.2.

```
[
    {
        "index": 13199985,
        "title": "NASA e SpaceX adiam lançamento da missão Europa Clipper devido ao
             furacão Milton",
        "url": "https://observador.pt/2024/10/07/nasa-e-spacex-adiam-lancamento-da-
            missao-europa-clipper-devido-ao-furacao-milton/",
        "pub_date": "2024-10-07 07:13:40",
        "feed_id": 23,
        "created_at": "2024-10-07 08:05:56.315153",
        "updated_at": "2024-10-07 08:05:56.315153",
        "twitter_shares": null,
```

```
        "facebook_shares": null,
        "summary": "Lançamento da missão Europa Clipper foi adiado devido ao furacã
            o Milton que vai afetar a Flórida, de onde partiria a nave. Estão
            previstas chuvas e ventos fortes no Cabo Canaveral.",
        "source_name": "Observador",
        "type_of_source": "national",
        "topics_BERTopic": "furacão, milton, missão, espacial, clipper",
        "topics_GSDMM": "Finfluencers, Líbano, Equador, Serralves, CMVM",
        "entities": "NASA, SpaceX, Milton, Flórida, Cabo Canaveral, Clipper",
        "types_of_entities": "ORG, ORG, PER, LOC, LOC, ACONTECIMENTO"
    }
]
```


Listing 5.2: Example entry from the `articles_everything` table.

Following, the connection to a data visualization tool that enables analysis and exploration of the data contained in the news headlines and summaries database is detailed.

5.3 Metabase

Metabase[1] is an open-source business intelligence and data visualization tool intended to allow users to ask questions about data, build dashboards, and generate reports, all through an easy-to-use web interface. Metabase offers a docker image that can be deployed locally and starts a Metabase container in any system that is running docker. Taking this into account, deploying metabase locally is an easy task to perform, as it is only necessary to run the following command:

```
docker run -d -p 3000:3000 --name metabase metabase/metabase
```

Having the application deployed locally, it was then necessary to connect to Metabase to get an authentication token, which was used to synchronize the PostgreSQL database with Metabase. This data visualization tool, on top of its abilities regarding visualization building and question asking about data, allows easy connection to PostgreSQL databases. However, it does not support connection to DuckDB databases, and that is exactly why it was crucial to transfer the data contained in the DuckDB tables back to the PostgreSQL database, as explained in the previous section.

Having all the data contained in the DuckDB tables transferred to the PostgreSQL database, the only thing left to do to finalize the process was to synchronize the PostgreSQL database with Metabase, using the token given by Metabase when connecting to it. Once all data were in the visualization tool, it was possible to take advantage of all features offered by it to better understand the data at hand.

In the next two sections, these Metabase features are used to ask questions and get a clearer picture of the news data. By using Metabase's tools, it is possible to better explore the information and understand important patterns and details in the dataset.

[1] https://www.metabase.com/

5.4 Publication Patterns

A comparison of the publication patterns between different sources is performed according to distinct temporal granularities: per month, per day of the month, per day of the week, and per hour. The ability to view multiple sources side by side on the same screen can offer valuable insights into their production cycles. Just for demonstration purposes, two sources were chosen to show this comparison: "Correio da Manhã" and "Record", the two national sources that published more news in 2024. These two sources are compared in terms of the percentage of their articles published during specific time periods, relative to the total number of articles published by them in 2024.

Looking first into coarser temporal granularities, we can observe in Figure 5.3 that "Record" published more news than "Correio da Manhã" in the first three months of the year, while "Correio da Manhã" was more active between April and October — with August being the exception, as in this month "Record" published more news than "Correio da Manhã". Observing now Figure 5.4, it can be concluded that both sources are similar in terms of days of the month when they prefer to publish news. Delving now into finer temporal granularities, in Figure 5.5, it can be observed that "Record" published more news on weekends than "Correio da Manhã", while the latter published more news during business days, with Thursday being an exception, when both news sources published almost the same quantity of news. Regarding the 24-hour cycle, we can observe in Figure 5.6 that, from 06:00 to 14:00 UTC, "Correio da Manhã" was more active than "Record", while the contrary occurs from 16:00 to 05:00 UTC, with 15:00 UTC being the turning point, when "Record" dominated in almost all hours in this time interval, with 00:00 UTC and 01:00 UTC being an exception.

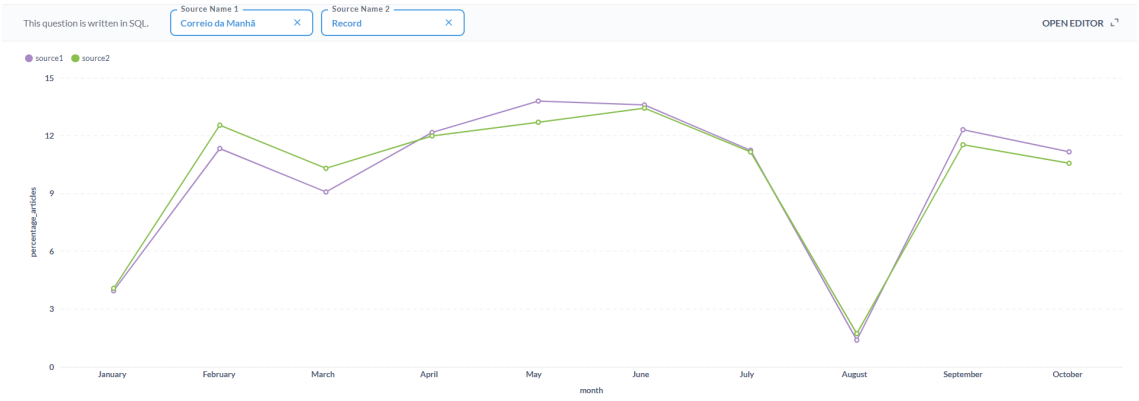

Figure 5.3: Articles published for each month of the year (%).

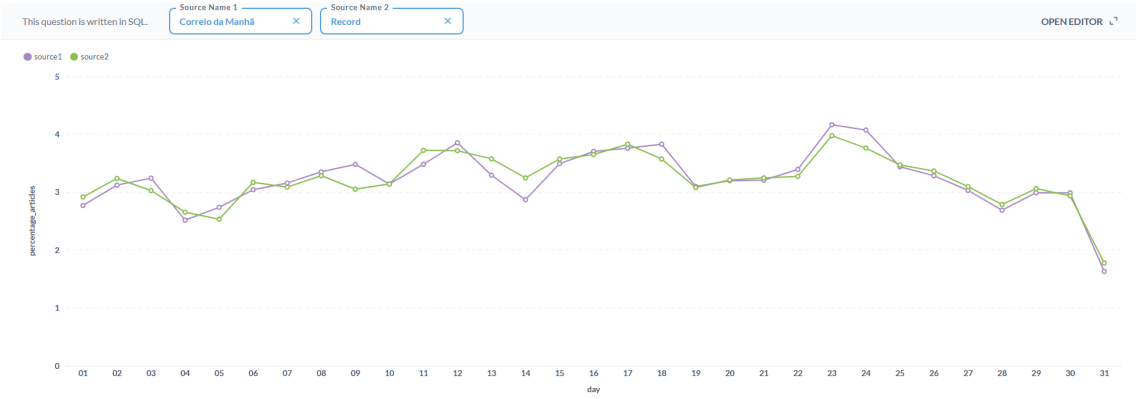

Figure 5.4: Articles published for each day of the month (%).

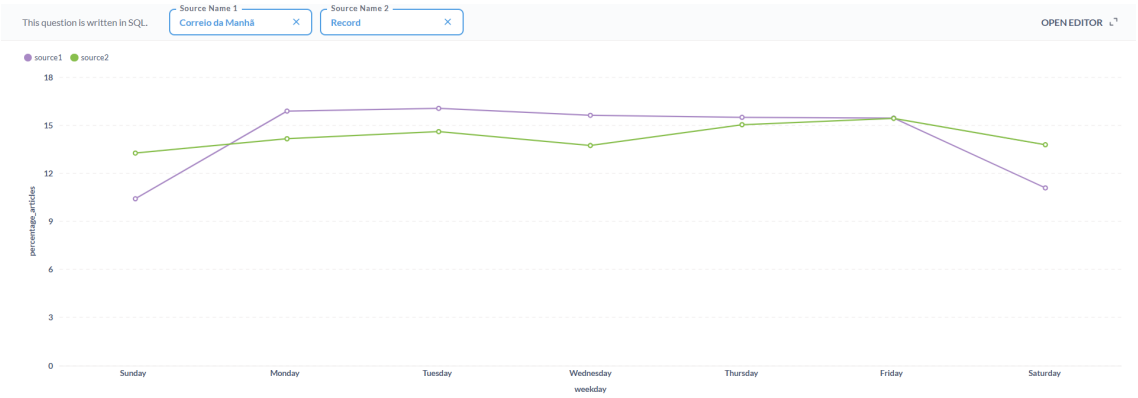

Figure 5.5: Articles published for each day of the week (%).

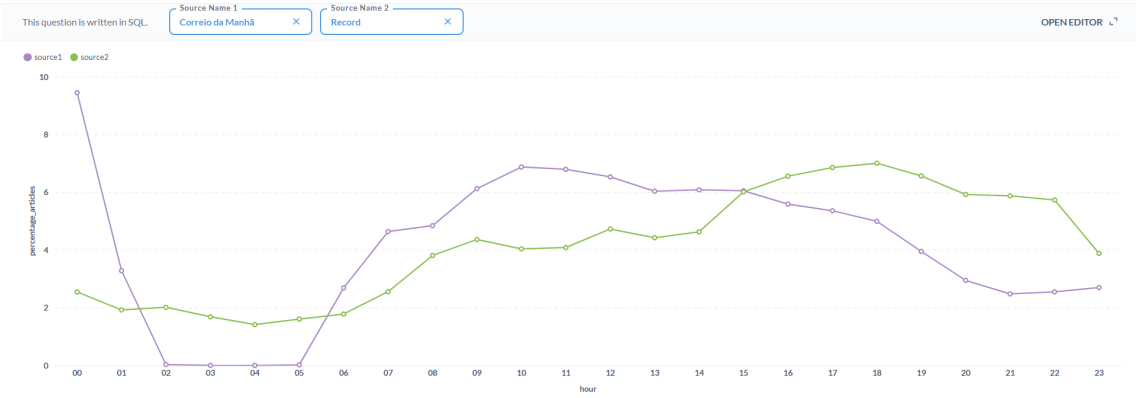

Figure 5.6: Articles published during a 24-hour cycle (%).

5.5 Variations in Coverage

Regarding variations in coverage, they can be observed by month, for the year 2024, across three different dimensions: topic, entity, and entity types. The ability to view multiple topics, entities,

or entity types side-by-side on the same screen can offer valuable insights into how much attention they receive.

For the case of entities, two individuals were selected for comparison: "Cristiano Ronaldo" and "Donald Trump", both of whom are frequently mentioned in the dataset — "Cristiano Ronaldo" is mentioned in 1356 news documents, and "Donald Trump" is mentioned in 1069 documents. Regarding entity types, the chosen ones were "PER" and "LOC", as these were the two entity types that received more mentions in 2024 (369,839 and 300,836 mentions, respectively). Finally, concerning topics, the chosen ones were "morreu, gregory, alf" and "irs, jovem" because they are related to some important events that took place in 2024. The percentages for entities and topics represent the share of articles that mention a given entity or topic out of the total number of articles published in the selected time period. Regarding the percentage for each entity type, it works in a different way, as it reflects how frequently that type appears in relation to all entity mentions within a specific time period. For example, if on August 12th there were 17 total entities mentioned across all articles published on that day — and 6 of those mentions were of type "PER" — then the percentage for "Person" on August 12th is (6 / 17) × 100 = 35.29%. This example can be seen in Figure 5.7. This percentage was not calculated based on the total number of articles, because a single article can mention multiple entities of the same type. Using article count as a base could lead to percentages exceeding 100%, which would be misleading. Therefore, the calculation was made relative to the total number of entity mentions.

	day text	all_entity_types text
1	12	LOC, PER, PER, LOC, LOC, ORG, LOC, PER, PER, LOC, LOC, TEMPO, PER, LOC, LOC, LOC, PER

Figure 5.7: Types of entities mentioned in articles published in August 12th.

5.5.1 Entity Coverage

Looking first at entity coverage per month, Figure 5.8 shows the coverage given to the two chosen entities in 2024 by Portuguese news sources. In June, mentions of Cristiano Ronaldo in Portuguese news increased sharply, going from about 0.3% of all articles to approximately 1.3%. This rise is in line with the start of Euro 2024, where Ronaldo was in the spotlight. Once Portugal's run in the tournament ended, the attention faded, and, from July onward, the attention given to Ronaldo decreased. Donald Trump, meanwhile, received very little coverage in the first half of 2024, which can be explained by the fact that the United States primary season did not attract much interest in Portugal. However, his numbers increased to more than 1.4% in July. Two major events explain this: first, the United States Supreme Court heard key arguments in one of Trump's legal cases; second, by mid-July, he had effectively secured the Republican nomination for the presidency. These developments brought a wave of news coverage in Portugal. After that, his media presence decreased again, settling around 0.5 – 0.7% in August and September.

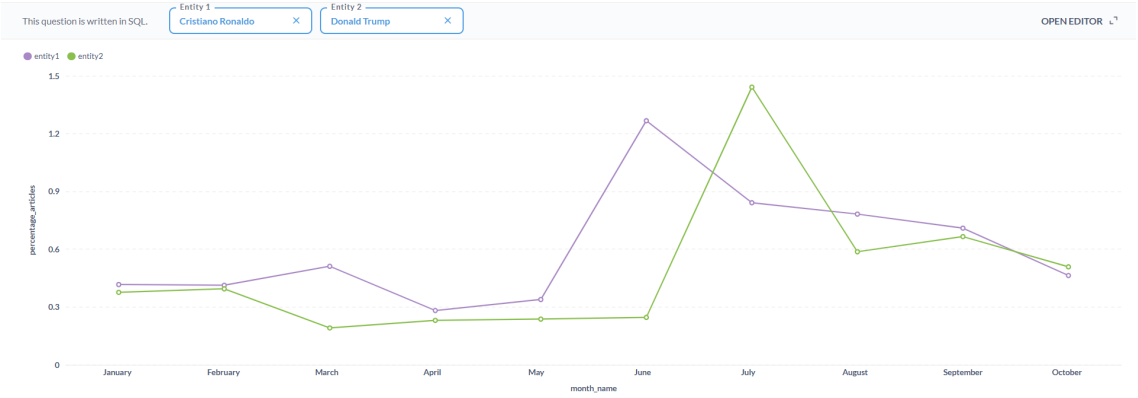

Figure 5.8: Percentage of articles mentioning "Cristiano Ronaldo" (represented by the purple line) and "Donald Trump" (represented by the green line) by month in 2024.

In terms of entity diversity, the sources were compared in terms of the number of entities each mentioned in relation to the total number of articles each one published. Figure 5.9 shows that "SAPO Desporto" exhibits the highest average number of entities mentioned per article, whereas "TSF" displays the lowest.

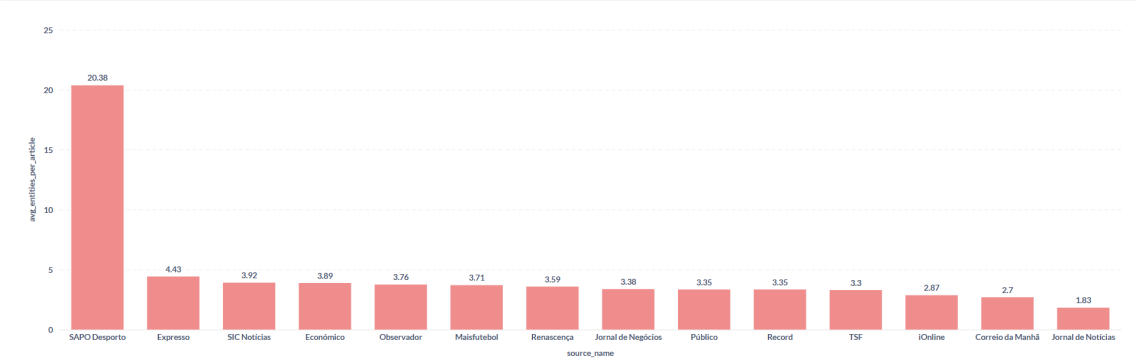

Figure 5.9: Average number of entities mentioned per article, per source.

Another analysis that can be done concerns the exploration of the most mentioned entities, both in general and per source. Figure 5.10 shows the five most mentioned entities in the dataset. "Portugal" appears as the most frequently mentioned entity, reflecting the domestic scope of the news content, and three of the five most mentioned entities are football clubs, which shows that Portuguese news sources focus a lot on football-related news. Looking now at Figure 5.11, which shows the most mentioned entities per source, considering only the five sources with the highest average number of entities mentioned per article in order to enhance the clarity and readability of the visualization, it is also unsurprising that football-related entities are the main focus of sports-dedicated sources, such as "SAPO Desporto", while receiving less emphasis in sources with a broader thematic scope. On the other hand, entities such as "Governo" or "PS" are the focus of more sources.

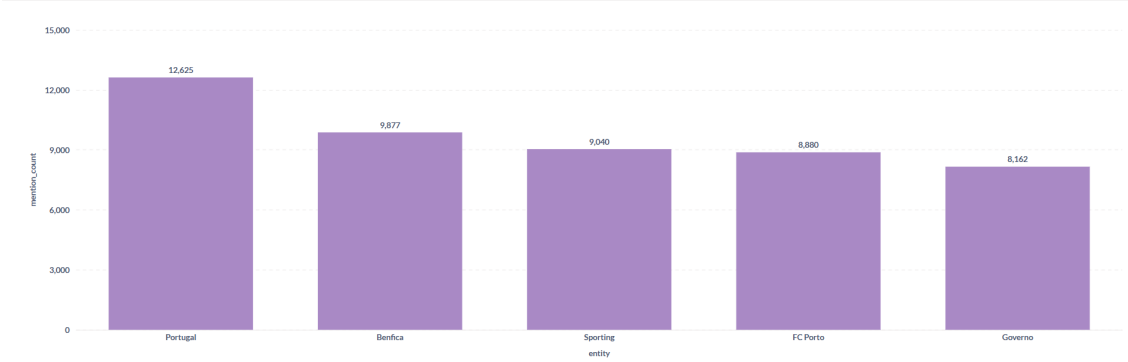

Figure 5.10: Most mentioned entities in the whole dataset.

Figure 5.11: Most mentioned entities per source, considering the five sources with the highest average number of entities mentioned per article.

5.5.2 Entity Type Coverage

Considering now entity type coverage, Figure 5.12 shows the coverage given to the two chosen entity types in 2024 by Portuguese news sources. "LOC" entities consistently receive more attention than "PER" entities across all months. This suggests a general editorial or topical focus more aligned with places than people, which shows that news sources' focus is more inclined toward geography (e.g., conflict zones, disasters, regional developments) than to people.

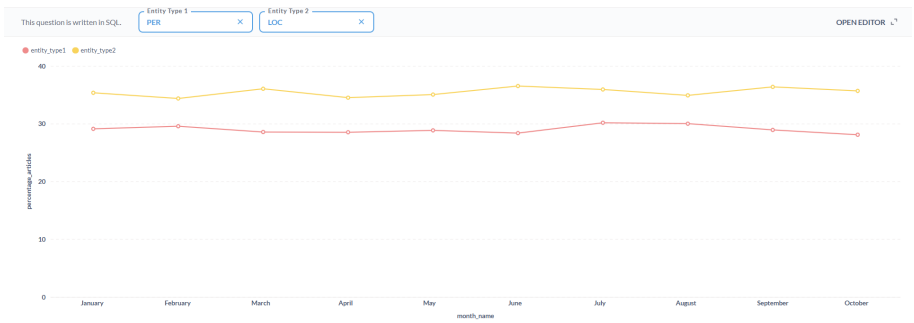

Figure 5.12: Percentage of articles mentioning "PER" entity types (represented by the red line) and "LOC" entity types (represented by the yellow line) by month in 2024.

It is also relevant to analyze the most frequently mentioned entity types, both in general and by source. Figure 5.13 presents the entity types with the highest number of mentions in the dataset. "Location" emerges as the most frequently referenced type, whereas "Work of Art" appears the least. This distribution suggests that news articles tend to prioritize geographic or geopolitical contexts over cultural or artistic references. The prevalence of location entities may reflect a focus on national and international events, while the limited presence of work-of-art entities indicates that cultural reporting occupies a relatively smaller portion of the news agenda. However, Figure 5.14 highlights that certain sources deviate from the overall trend in terms of which entity types are most frequently mentioned. Three examples of this are "Maisfutebol", "SAPO Desporto", and "Record", three sports-related news sources, which mention more Person entity types than Location entity types. This pattern is likely due to their editorial focus on individual athletes, whose personal stories and achievements are central to sports reporting. Another two examples of this deviation are "Económico" and "Jornal de Negócios", both of which focus primarily on economic and financial news. Although, in general, Organization entities receive fewer mentions than Person entities, in these sources the opposite trend is observed, with organizations being mentioned more frequently. This can be attributed to the nature of their content, which places greater emphasis on companies and institutions rather than individual figures. These deviations underscore how the thematic orientation of a news source shapes the type of entities it mentions.

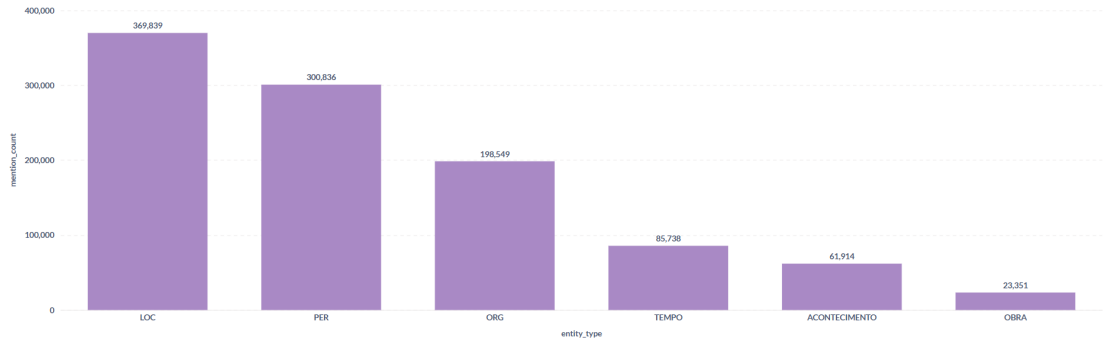

Figure 5.13: Entity types ordered by number of mentions in the whole dataset.

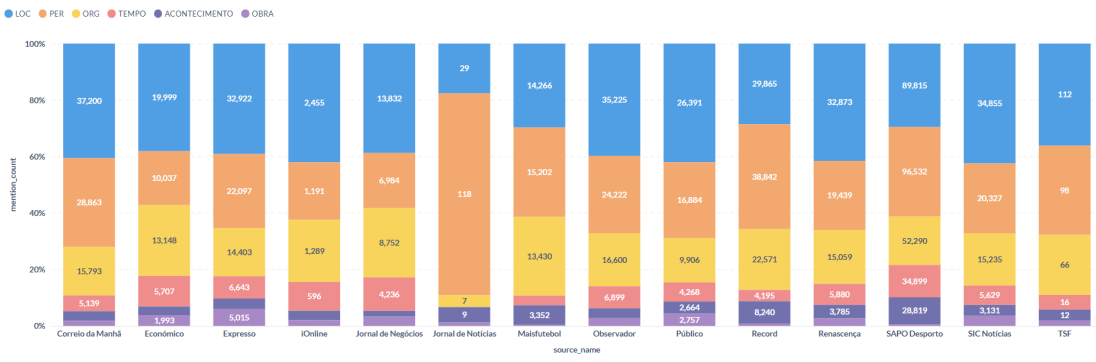

Figure 5.14: Entity types ordered by number of mentions per source.

5.5.3 Topic Coverage

Finally, with respect to topic coverage, two distinct conclusions were reached regarding the most effective topic modeling tool and topic labeling strategy. The user survey indicated that the optimal combination was GSDMM for topic modeling and Yake! for topic labeling. In contrast, the manual evaluation suggested that BERTopic, along with its automatically generated topic keywords, provided the most accurate and coherent representation for topic extraction. Figures 5.15 and 5.16 present the number of topics identified by each Portuguese news source in 2024 using BERTopic and GSDMM, respectively. The results indicate no significant disparity between the two models in terms of topic distribution across sources: BERTopic identified more topics in seven of the 14 sources, while GSDMM did so for the remaining seven.

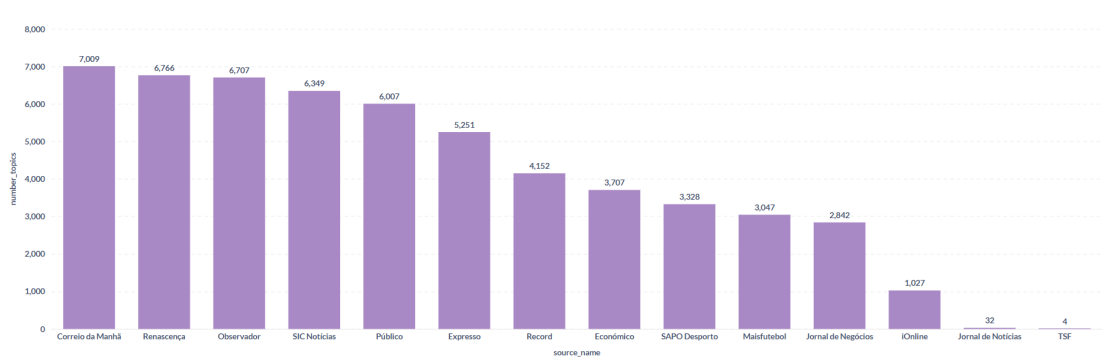

Figure 5.15: Number of topics covered per source by BERTopic.

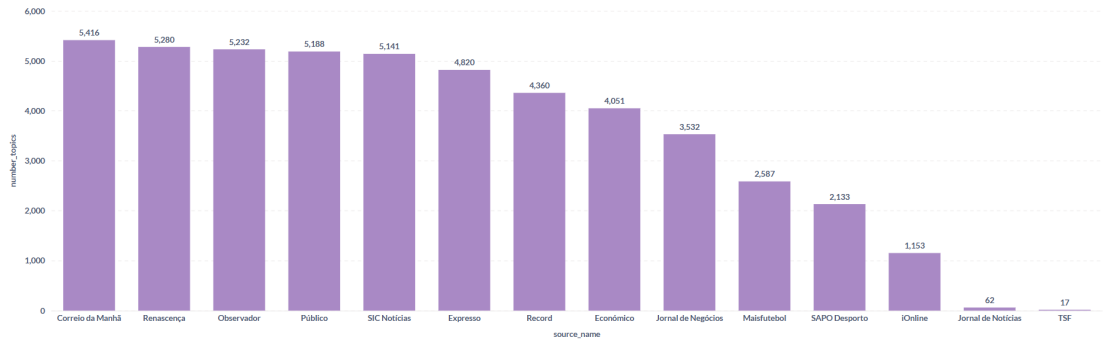

Figure 5.16: Number of topics covered per source by GSDMM.

Regarding topic diversity, some differences can be seen between the results given by the two topic modeling tools. When comparing the number of topics covered by each source with the total number of articles posted by them, it can be seen in Figure 5.17 that "iOnline" exhibits the highest number of topics in relation to the number of articles published, while "TSF" displays the lowest for the case of BERTopic. In the case of "iOnline", there is one new topic per each two news articles, while in the case of "TSF" there is one new topic per each 14 news articles. On the other hand, for the case of GSDMM, in Figure 5.18, it can be observed that "Jornal de Notícias" displays

the highest number of topics in relation to the number of articles published, while "Record" shows the lowest.

Another interesting observation concerns the expectations regarding the news sources exhibiting the highest and lowest average number of topics per article. As shown in Figure 5.17, in the case of BERTopic news sources with a wider thematic scope such as "iOnline", "Jornal de Notícias", "Público", "Renascença", and "Sic Notícias" display the highest average number of topics per article, while more thematically focused news sources such as "Económico", "Jornal de Negócios", "Maisfutebol", "SAPO Desporto" and "Record" demonstrate the lowest. Correio da Manhã" and "TSF" — mainly "TSF" — deviate the most from expectations, as they are two news sources with wider thematic scope, yet rank among those with the lowest average number of topics per article. In the case of GSDMM, "Jornal de Negócios" and "Económico" are the biggest outliers, as they present a higher average number of topics per article than news sources with more diverse editorial content. "Correio da Manhã" also slightly deviates from the expectations here.

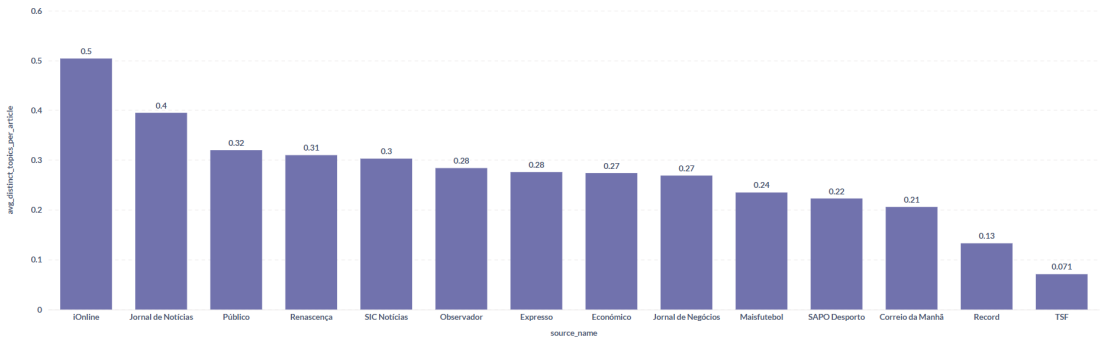

Figure 5.17: Average number of topics covered per article, per source considering BERTopic modeling.

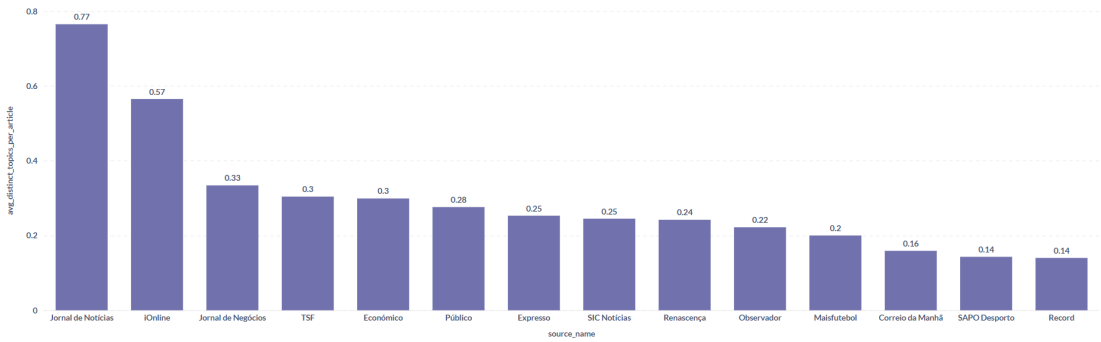

Figure 5.18: Average number of topics covered per article, per source, considering GSDMM modeling.

To further analyze topic coverage, Figures 5.19 and 5.20 illustrate the monthly percentage of articles assigned to two selected topics in 2024, as extracted by BERTopic and GSDMM, respectively. These visualizations highlight how each model captures the evolution and prevalence of

these two topics over time. When comparing the two figures, it is evident that the topic "morreu, gregory, alf" received greater coverage in GSDMM than in BERTopic. Since both models were applied to the same set of news documents, this discrepancy may suggest that either BERTopic failed to assign some relevant articles to this topic, or that GSDMM included articles that were not truly representative of it. Looking at the topic "irs, jovem", regarding BERTopic it received increasing attention from August to October, while in GSDMM it received a peak of attention in July. This also highlights a discrepancy in how the two topic modeling tools assign topics to articles.

Another interesting observation can be made by confronting the results with a study titled "The Lifespan of News Stories" [24]. Although that study measures the attention consumers give to a particular topic or event, a parallel can be drawn with the present analysis of topic coverage, where attention is assessed not from the perspective of the audience, but rather based on the prominence that media outlets assign to specific topics. The topics "morreu, gregory, alf" and "irs, jovem" in the cases of BERTopic and GSDMM are symmetric topics, characterized by a steady climb and fall [24].

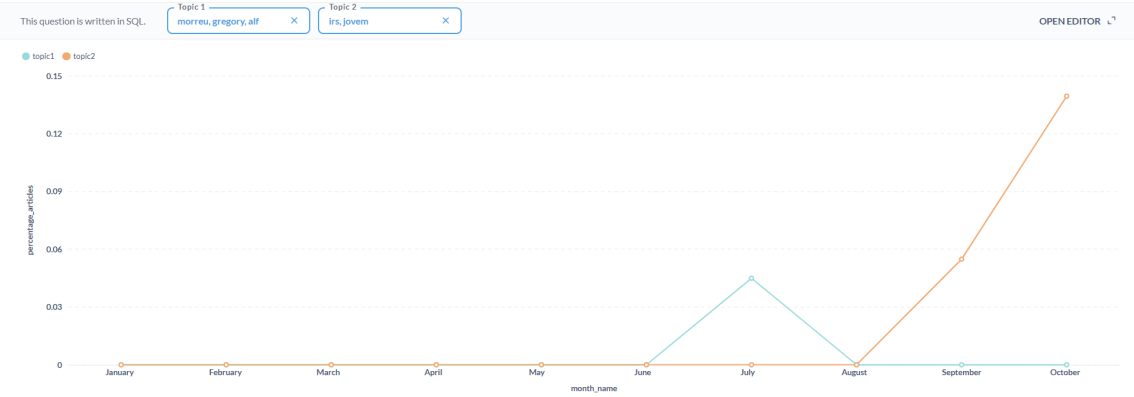

Figure 5.19: Percentage of articles mentioning "morreu, gregory, alf" (represented by the blue line) and "irs, jovem" (represented by the orange line) by month in 2024 in the case of BERTopic.

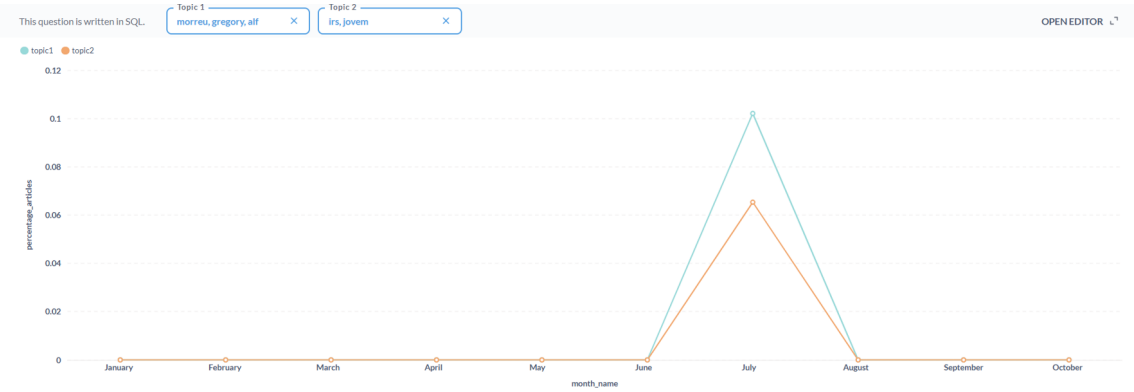

Figure 5.20: Percentage of articles mentioning "morreu, gregory, alf" (represented by the blue line) and "irs, jovem" (represented by the orange line) by month in 2024 in the case of GSDMM.

Similarly to the cases of entity and entity type coverage, it is also of interest to analyze the most prominent topics in the news documents present in the dataset. First, considering the case of BERTopic, Figure 5.21 shows that the five most prominent topics in the entire dataset revolve around football, which is consistent with the observation that three of the five most frequently mentioned entities are also football-related. Delving now into the most covered topics per source, Figure 5.22 shows the most covered topics per source, this time considering only the three sources with the highest average number of topics mentioned per article, plus two other more thematically specific sources for comparison purposes, also in order to enhance the clarity and readability of the visualization. It can be seen that "Record" focuses mainly on football-related topics and "Jornal de Negócios" on topics related to economy, as expected. The remaining three sources align with expectations by covering a broader range of thematically distinct topics. Delving into the case of GSDMM, Figure 5.23 shows that the five most prominent topics in the entire dataset revolve around sports in general. Looking at Figure 5.24, which considers the same news sources selected in the case of BERTopic, it can also be concluded that the topic distributions for each source in the case of GSDMM generally align well with the known editorial profiles of the sources, reflecting their thematic focus.

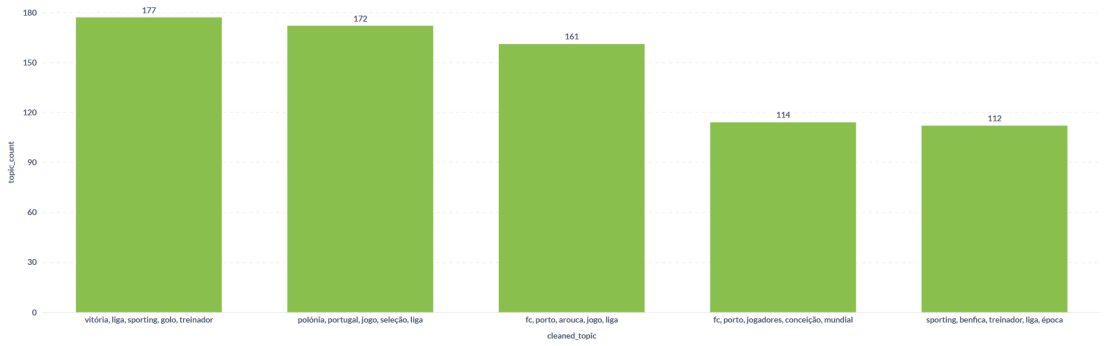

Figure 5.21: Most covered topics in the whole dataset in the case of BERTopic.

Figure 5.22: Most covered topics per source in the case of BERTopic, considering the three sources with the highest average number of topics mentioned per article and two other sources.

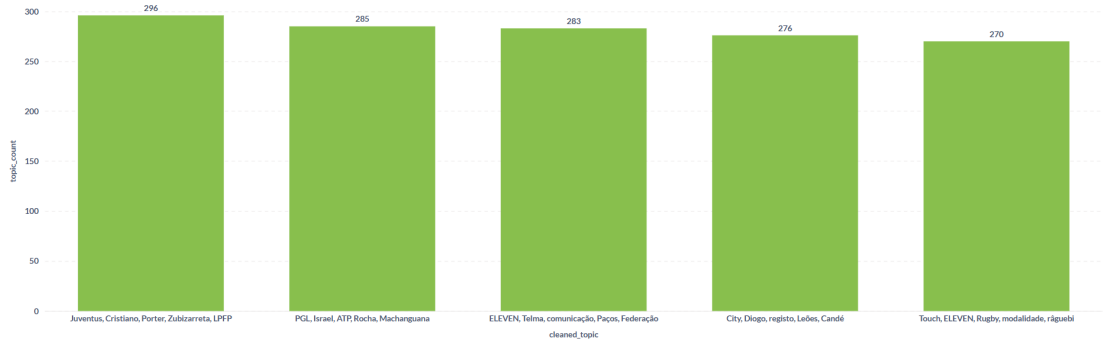

Figure 5.23: Most covered topics in the whole dataset in the case of GSDMM.

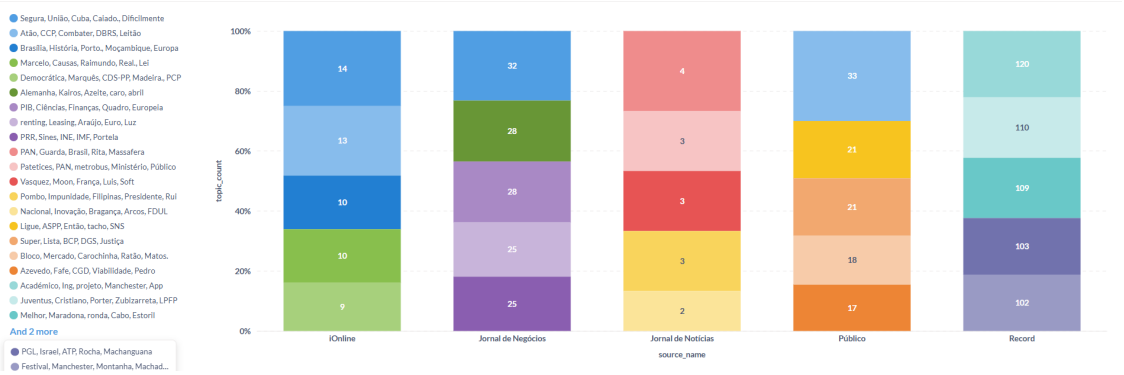

Figure 5.24: Most covered topics per source in the case of GSDMM, considering the three sources with the highest average number of topics mentioned per article and two other sources.

5.6 Summary

Chapter 5 presents and discusses the outcomes obtained from applying the methodologies described in the previous chapter. The goal of this section was to evaluate the effectiveness of the developed pipeline in extracting topics and entities from Portuguese news content, as well as to analyze the enriched dataset in terms of publication patterns, variations in coverage, and diversity.

The chapter begins by detailing the results of the manual evaluations and user surveys conducted to assess the performance of the topic modeling and NER tools. For the entity extraction task, the spaCy multilingual model achieved the highest F1-score in the detection of core entities such as persons, organizations, and locations, while the BERTimbau-based model showed superior performance in recognizing extended entity types such as events, time expressions, and works of art. Based on this complementary performance, a hybrid setup was adopted in which spaCy was used for core entities and BERTimbau for the extended ones. The evaluation of topic modeling tools followed a similar dual approach. While BERTopic demonstrated the highest hit rate in a manual evaluation —successfully identifying relevant topics in 64.39% of the cases — GSDMM was preferred by the user survey participants, obtaining the highest average rating, followed by

BERTopic and LDA. These findings suggest that different topic modeling strategies may excel under different criteria.

After selecting the most adequate tools, the dataset was enriched with the extracted topics and entities, resulting in a structured corpus with semantic annotations. This enriched dataset was then visualized using Metabase, which enabled the exploration of publication patterns and coverage trends through an interactive dashboard. Metabase proved useful for analyzing temporal trends across sources and for inspecting entity and topic occurrences over time and across sources.

Overall, this chapter validates the pipeline developed and demonstrates the value of combining different informatics fields to study the Portuguese online news landscape.

Chapter 6

Conclusions and Future Work

6.1 Final Remarks

This dissertation addressed the challenge of analyzing a large-scale dataset of Portuguese online news with the primary objective of building a solid understanding of the structure, content, and dynamics of online news articles published in Portugal. The initial task involved applying NLP techniques to extract and enrich the data with relevant topics and named entities. In addition to the semantic enrichment of the dataset, the work also focused on the analysis of publication patterns, aiming to uncover how different sources publish content over time. This included identifying variations in publication frequency across different temporal granularities — per month, per day of the month, per day of the week, and per hour. Together, these analyses offer a deeper understanding of not only what is being published, but also when and by whom, contributing to a more comprehensive view of the Portuguese digital news ecosystem.

The first part of the work involved a systematic literature review aimed at understanding the current state of research in the fields of topic and entity extraction, news publication patterns, and media diversity. This review helped in identifying the most relevant methodologies, tools, and evaluation strategies used in previous studies, as well as in highlighting existing gaps and challenges. It provided a solid foundation for defining the scope of the project and guided the selection of appropriate techniques for the subsequent experimental phase.

Building upon the literature review, the state of the art presented a more detailed overview of current techniques and developments in variations in coverage, news publication patterns, and media diversity. This chapter offered a critical examination of widely used tools and methods, highlighting both their capabilities and limitations. It also helped frame the technological and academic context of the dissertation, showing how current research trends align with or diverge from the goals of this work. Key insights from this chapter helped guide the choice of tools and techniques later implemented in the experimental phase.

The implemented solution included data cleaning, corpus characterization, and the application of different tools for topic modeling, NER, and different strategies for topic labeling. The evaluation of these tools and strategies was conducted through both analytical assessments and a user

survey, allowing a practical comparison between different approaches. These enriched outputs were then visualized using Metabase, enabling the observation of content and publication patterns across different sources and temporal granularities.

The results confirm that combining structured topic extraction and entity recognition methods allows for a deeper understanding of news content and publication behaviors. Moreover, the enriched dataset supports further exploration of trends and coverage variation in Portuguese media. This work thus contributes to the broader field of news analysis and offers a framework for future academic studies and media-related applications.

Drawing on the work carried out, the following responses are provided to address the research questions:

RQ1: What is the best methodology to follow to extract topics from news documents?

To determine the best methodology for topic extraction from Portuguese news documents, the study evaluated three topic modeling tools: LDA, BERTopic, and GSDMM. A two-phase evaluation was conducted. The first included a manual evaluation, where news from October 25th was clustered using each model, and a manual comparison was made between the cluster's top keywords and the actual content of the articles. BERTopic achieved the highest hit rate (64.39%), followed by GSDMM (58.50%) and LDA (54.47%). The second included a user survey, where 20 participants rated how well each news cluster (produced by each model) represented a coherent topic. According to the weighted mean scores from this evaluation, GSDMM outperformed both BERTopic and LDA, scoring 4.49, compared to 4.09 for BERTopic and 3.61 for LDA. Given this evidence, two different conclusions were reached: on the one hand, according to the manual evaluation, BERTopic was the most adequate topic modeling tool, and the words given by it represented better each topic than the keywords given by Yake! On the other hand, the user survey revealed that GSDMM was the most adequate topic modeling tool, while the keywords extracted with Yake! worked better as labels for the topics than the word representations given by GSDMM. These results were further compared and discussed in Section 5.5 of Chapter 5.

RQ2: What is the best methodology to follow to extract entities from news documents?

Regarding entity extraction, the study evaluated three tools: spaCy, GLiNER, and a BERTimbau-based model. The evaluation was based on the number of entities correctly identified and labeled (True Positives), entities incorrectly labeled or non-entities wrongly identified (False Positives), and missed entities that should have been detected (False Negatives). To determine the most effective tool, the F1-score was calculated, balancing both precision and recall. The spaCy multilingual model was determined to be the most robust option to extract entities of type Person, Location, and Organization, reaching an F1-score of 77.89% for these types of entities, while the BERTimbau-based model was considered the most adequate to extract entities of type Event, Work of Art, and Time, as it achieved a better score than GLiNER, the other neural model capable of identifying these types of entities — 71.43% of BERTimbau against 64.41% of GLiNER.

RQ3: How can the publication patterns of different news sources be analyzed through the volume and frequency of news articles published by them?

Publication patterns can be analyzed by aggregating the volume and temporal distribution of articles across various sources. In this study, the data were first organized by source and timestamp and then visualized using Metabase, which enabled interactive dashboards and charts. Temporal granularities such as daily, weekly, and monthly distributions were examined to uncover trends in news production. This allowed for comparisons in output intensity, identification of publication peaks, and detection of content gaps. These patterns reflect editorial routines and operational characteristics of each outlet, providing insight into the structural behavior of Portuguese news media throughout 2024.

6.2 Future Work

Although all research questions were answered, there is still room for improvement in this work. First, as explained in Chapter 4, only news published by Portuguese news sources in 2024 was utilized in this study, for simplicity and due to time constraints. One potential improvement to this study would be to incorporate the full corpus of news articles published by Portuguese news sources, rather than limiting the analysis to those from the year 2024. Expanding the temporal scope of the dataset would enable a more comprehensive examination of publication patterns and topic and entity coverage over time. Such an approach would enable longitudinal comparisons across multiple years, offering deeper insights into trends, shifts in editorial focus, and the evolution of journalistic practices within the Portuguese news ecosystem.

An additional enhancement could involve a more comprehensive treatment of the concept of diversity. While this topic was briefly addressed in Chapter 5 — particularly through the analysis of the number of topics and entities per source — it was examined primarily through a limited lens. As outlined in Chapter 3, diversity can be assessed across three key dimensions: "variety", "balance", and "disparity". However, the current analysis focuses slightly and exclusively on "variety", considering only the count of distinct entities or topics. The dimensions of "balance" — which relates to the evenness of distribution — and "disparity" — which concerns the degree of difference between categories — were not explored. Therefore, a valuable direction for improvement would be to incorporate these additional dimensions into the measurement of topic and entity diversity, enabling a more nuanced and multidimensional understanding of the dataset.

Another potential refinement concerns the methodology for topic labeling. As discussed in Chapter 4, two main approaches were explored: utilizing the word distributions provided by topic modeling tools, and extracting representative keywords from news documents using the YAKE! keyword extractor. While both methods offer insight into the content of each topic, they primarily generate a set of indicative terms rather than assigning a concise, descriptive label. This limits the interpretability of the resulting topics, particularly in contexts where clear, human-readable labels are essential for downstream analysis or visualization. An avenue for further development would be the design of a dedicated labeling strategy capable of assigning a specific, coherent label to each

topic. This could involve leveraging external knowledge bases or even employing large language models to generate semantically meaningful labels that summarize the underlying theme of each topic more effectively.

Finally, the development of a web application integrated with Metabase could be a valuable addition. This application would be inspired by the MediaCloud platform[1] and would serve as an interactive interface for exploring the dataset of news headlines and summaries in a comprehensive manner. It would enable users to examine patterns of publication across all news sources included in the dataset, as well as variations in coverage across entities and topics. Such exploration could be conducted across multiple dimensions, including different temporal granularities, media sources, and types of sources. By offering this level of granularity and flexibility, the tool would facilitate a deeper understanding of the Portuguese online news ecosystem. This application would be especially valuable for media professionals, researchers, and academics, providing them with a powerful resource to investigate editorial practices, media attention cycles, and content diversity in a transparent and data-driven manner.

[1] https://search.mediacloud.org/

References

[1] Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131, 2023.

[2] Wan Najmiyyah Wan Md Adnan and Sarimah Shamsudin. A systematic approach to identifying social actors in online news corpora. In *Proceedings of the 7th International Conference on Education and Multimedia Technology, ICEMT 2023, Tokyo, Japan, August 29-31, 2023*, pages 383–390. ACM, 2023.

[3] Alan B. Albarran. *Media Economics*. Iowa State Press, Ames, 2nd edition, 2002.

[4] Eran Amsalem, Yair Fogel-Dror, Shaul R. Shenhav, and Tamir Sheafer. Fine-Grained Analysis of Diversity Levels in the News. *Communication Methods and Measures*, 14(4):266 – 284, 2020.

[5] Jisun An, Daniele Quercia, Meeyoung Cha, Krishna P. Gummadi, and Jon Crowcroft. Sharing political news: the balancing act of intimacy and socialization in selective exposure. *EPJ Data Sci.*, 3(1):12, 2014.

[6] Evangelia Avraam, Andreas Veglis, and Charalampos Dimoulas. Publishing Patterns in Greek Media Websites. *Social Sciences*, 10(2):59, February 2021.

[7] Olusola Babalola, Bolanle Ojokoh, and Olutayo Boyinbode. Comprehensive evaluation of LDA, NMF, and BERTopic's performance on news headline topic modeling. *Journal of Computing Theories and Applications*, 2(2):268–289, 2024.

[8] Kathleen Beckers, Andrea Masini, Julie Sevenans, Miriam van der Burg, Julie De Smedt, Hilde Van den Bulck, and Stefaan Walgrave. Are newspapers' news stories becoming more alike? Media content diversity in Belgium, 1983–2013. *Journalism*, 20(12):1665 – 1683, 2019.

[9] Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774, October 2018.

[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[11] Jan Boehmer, Serena Carpenter, and Frederick Fico. More of the Same?: Influences on source use and source affiliation diversity in for-profit and nonprofit news organizational content. *Journalism Studies*, 20(2):173 – 192, 2019.

[12] Engin Bozdag, Qi Gao, Geert-Jan Houben, and Martijn Warnier. Does offline political segregation affect the filter bubble? an empirical analysis of information diversity for dutch and turkish twitter users. *Comput. Hum. Behav.*, 41:405–415, 2014.

[13] Maria Carla Calzarossa and Daniele Tessera. Time series analysis of the dynamics of news websites. In *Parallel and Distributed Computing, Applications and Technologies, PDCAT Proceedings*, pages 529 – 533, 2012.

[14] Maria Carla Calzarossa and Daniele Tessera. Analysis and forecasting of web content dynamics. In Leonard Barolli, Makoto Takizawa, Tomoya Enokido, Marek R. Ogiela, Lidia Ogiela, and Nadeem Javaid, editors, *32nd International Conference on Advanced Information Networking and Applications Workshops, AINA 2018 workshops, Krakow, Poland, May 16-18, 2018*, pages 12–17. IEEE Computer Society, 2018.

[15] Erik Cambria and Bebo White. Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*, 9(2):48–57, May 2014.

[16] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Inf. Sci.*, 509:257–289, 2020.

[17] Giuseppe Carrino, Angelo Di Iorio, and Gioele Barabucci. Comparison of news commonality and churn in international news outlets with TARO. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media (HT '23)*, pages 1–10. Association for Computing Machinery, 2023.

[18] Jiann-Fuh Chen, Wei-Ming Wang, and Chao-Ming Huang. Analysis of an adaptive time-series autoregressive moving-average (arma) model for short-term load forecasting. *Electric Power Systems Research*, 34(3):187–196, 1995.

[19] Xi Chen, Scott A. Hale, David Jurgens, Mattia Samory, Ethan Zuckerman, and Przemyslaw A. Grabowicz. Global News Synchrony and Diversity During the Start of the COVID-19 Pandemic. In *WWW 2024 - Proceedings of the ACM Web Conference*, pages 2639 – 2650, 2024.

[20] Xu Chen, Judith Gelernter, Han Zhang, and Jin Liu. Multi-lingual geoparsing based on machine translation. *Future Generation Computer Systems*, 96:667–677, July 2019.

[21] Cédric Courtois, Laura Slechten, and Lennert Coenen. Challenging google search filter bubbles in social and political information: Disconforming evidence from a digital methods case study. *Telematics Informatics*, 35(7):2006–2015, 2018.

[22] Ernesto de León and Susan Vermeer. The News Sharing Gap: Divergence in Online Political News Publication and Dissemination Patterns across Elections and Countries. *Digital Journalism*, 11(2):343 – 362, 2023.

[23] Giuseppe De Martino and Serena Spina. Exploiting the time-dynamics of news diffusion on the Internet through a generalized Susceptible-Infected model. *Physica A: Statistical Mechanics and its Applications*, 438:634 – 644, 2015.

[24] Schema Design and Google Trends. The lifespan of news stories, 2019. Available at https://www.newslifespan.com/, last accessed in October 2024.

[25] T. Devezas, J. Devezas, and S. Nunes. Exploring a large news collection using visualization tools. In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016*, volume 1568, pages 48 – 53, 2016.

[26] Tiago Devezas, Sérgio Nunes, and Manuel Rodríguez. MediaViz: An Interactive Visualization Platform for Online Media Studies. In *Proceeding of the 2015 International Workshop on Human-centric Independent Computing*, pages 7–11, 2015.

[27] Burak Doğu. Turkey's news media landscape in Twitter: Mapping interconnections among diversity. *Journalism*, 21(5):688 – 706, 2020.

[28] Jennifer Ecker. Labeling results of topic models: Word sense disambiguation as key method for automatic topic labeling with germanet. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10014–10022, Torino, Italia, 2024. ELRA and ICCL.

[29] Erick Elejalde, Leo Ferres, Eelco Herder, and Johan Bollen. Quantifying the ecological diversity and health of online news. *Journal of Computational Science*, 27:218 – 226, 2018.

[30] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S. Yu, and Hongjun Lu. Parameter free bursty events detection in text streams. In Klemens Böhm, Christian S. Jensen, Laura M. Haas, Martin L. Kersten, Per-Åke Larson, and Beng Chin Ooi, editors, *Proceedings of the 31st International Conference on Very Large Data Bases, Trondheim, Norway, August 30 - September 2, 2005*, pages 181–192. ACM, 2005.

[31] G.Box, G.Jenkins, and G.Reinsel. *Time Series Analysis: Forecasting and Control*. Palgrave Macmillan UK, 2008.

[32] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[33] Lei Guo. Media Agenda Diversity and Intermedia Agenda Setting in a Controlled Media Environment: A Computational Analysis of China's Online News. *Journalism Studies*, 20(16):2460 – 2477, 2019.

[34] Lei Guo, Chris J. Vargo, Zixuan Pan, Weicong Ding, and Prakash Ishwar. Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling. *Journalism & Mass Communication Quarterly*, 93(2):332–359, June 2016.

[35] Ahmed Hamdi, Elvys Linhares Pontes, and Antoine Doucet. Annotation guidelines for named entity recognition, entity linking and stance detection. Zenodo, March 2021. Accessed 2025-06-19.

[36] J.D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[37] Takako Hashimoto, David Lawrence Shepard, Tetsuji Kuboyama, Kilho Shin, Ryota Kobayashi, and Takeaki Uno. Analyzing temporal patterns of topic diversity using graph clustering. *The Journal of Supercomputing*, 77(5):4375–4388, May 2021.

[38] Natali Helberger. Merely Facilitating or Actively Stimulating: Diverse Media Choices? Public Service Media at the Crossroad. *International Journal of Communication*, 9:1324–1340, January 2015.

[39] Jonathan Hendrickx, Pieter Ballon, and Heritiana Ranaivoson. Dissecting news diversity: An integrated conceptual framework. *Journalism*, 23(8):1751 – 1769, 2022.

[40] Ian Holmes, Keith Harris, and Christopher Quince. Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLOS ONE*, 7(2):e30126, 2012.

[41] Edda Humprecht and Florin Büchel. More of the Same or Marketplace of Opinions? A Cross-National Comparison of Diversity in Online News Reporting. *International Journal of Press/Politics*, 18(4):436 – 461, 2013.

[42] Edda Humprecht and Frank Esser. Diversity in Online News: On the importance of ownership types and media system types. *Journalism Studies*, 19(12):1825 – 1847, 2018.

[43] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In Keh-Jiann Chen, Chu-Ren Huang, and Richard Sproat, editors, *Proceedings of the 10th Research on Computational Linguistics International Conference*, pages 19–33, Taipei, Taiwan, August 1997. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

[44] Glen Joris, Frederik De Grove, Kristin Van Damme, and Lieven De Marez. News Diversity Reconsidered: A Systematic Literature Review Unraveling the Diversity in Conceptualizations. *Journalism Studies*, 21(13):1893 – 1912, 2020.

[45] Keisuke Kiritoshi and Qiang Ma. A diversity-seeking mobile news app based on difference analysis of news articles. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9262:73 – 81, 2015.

[46] Laura Koivunen-Niemi and Masood Masoodian. Visualizing narrative patterns in online news media. *Multimedia Tools and Applications*, 79(1-2):919 – 946, 2020. Type: Article.

[47] Kean Shern Kok and Hui Na Chua. Using Word2Vec-LDA-Cosine Similarity for Discovering News Dissemination Pattern to Support Government–Citizen Engagement. *Lecture Notes in Networks and Systems*, 288:703 – 716, 2022.

[48] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1536–1545, Portland, Oregon, 2011. Association for Computational Linguistics.

[49] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. Event registry: Learning about world events from news. *arXiv preprint arXiv:1504.01433*, 2014.

[50] Lu Li. Data news dissemination strategy for decision making using new media platform. *Soft Computing*, 26(20):10677 – 10685, 2022.

[51] Fen Lin and Xinzhi Zhang. Movement-press dynamics and news diffusion: A typology of activism in digital china*. *China Review*, 18(2):33 – 63, 2018.

[52] Nairong Liu, Haizhong An, Xiangyun Gao, Huajiao Li, and Xiaoqing Hao. Breaking news dissemination in the media via propagation behavior based on complex network theory. *Physica A: Statistical Mechanics and its Applications*, 453:44 – 54, 2016.

[53] Ping Liu, Karthik Shivaram, Aron Culotta, Matthew A. Shapiro, and Mustafa Bilgic. The interaction between political typology and filter bubbles in news recommendation algorithms. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, editors, *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 3791–3801. ACM / IW3C2, 2021.

[54] Felicia Loecherbach, Judith Moeller, Damian Trilling, and Wouter Van Atteveldt. The Unified Framework of Media Diversity: A Systematic Literature Review. *Digital Journalism*, 8(5):605–642, May 2020.

[55] Yue Lu, Qiaozhu Mei, and ChengXiang Zhai. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Inf. Retr.*, 14(2):178–203, 2011.

[56] M.D. Madhushika, Supunmali Ahangama, and D.A. Rajapaksha. Analyzing the Impact of Social Media on Sinhala News Dissemination in Mass Media. In *ICARC 2022 - 2nd International Conference on Advanced Research in Computing: Towards a Digitally Empowered Society*, pages 177 – 182, 2022.

[57] Melanie Magin, Birgit Stark, Olaf Jandura, Linards Udris, Andreas Riedl, Miriam Klein, Mark Eisenegger, Raphael Kösters, and Brigitte Hofstetter Furrer. Seeing the Whole Picture. Towards a Multi-perspective Approach to News Content Diversity based on Liberal and Deliberative Models of Democracy. *Journalism Studies*, 24(5):669 – 696, 2023.

[58] Kazuhisa Makino and Takeaki Uno. New algorithms for enumerating all maximal cliques. In Torben Hagerup and Jyrki Katajainen, editors, *Algorithm Theory - SWAT 2004, 9th Scandinavian Workshop on Algorithm Theory, Humlebaek, Denmark, July 8-10, 2004, Proceedings*, volume 3111 of *Lecture Notes in Computer Science*, pages 260–272. Springer, 2004.

[59] Andrea Masini, Peter Van Aelst, Thomas Zerback, Carsten Reinemann, Paolo Mancini, Marco Mazzoni, Marco Damiani, and Sharon Coen. Measuring and Explaining the Diversity of Voices and Viewpoints in the News: A comparative study on the determinants of content diversity of immigration news. *Journalism Studies*, 19(15):2324 – 2343, 2018.

[60] Andrew K. McCallum. MALLET: A Machine Learning for Language Toolkit, 2002.

[61] Daniel G. Mcdonald and John Dimmick. The Conceptualization and Measurement of Diversity. *Communication Research*, 30(1):60–79, February 2003.

[62] Denis McQuail. Media Performance. In Wolfgang Donsbach, editor, *The International Encyclopedia of Communication*. Wiley, 1 edition, June 1992.

[63] Ida Mele, Seyed Ali Bahrainian, and Fabio Crestani. Linking News across Multiple Streams for Timeliness Analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 767 – 776, Singapore Singapore, November 2017. ACM.

[64] Carlos Mendez, Fernando Mendez, Vasiliki Triga, and Juan Miguel Carrascosa. EU Cohesion Policy under the Media Spotlight: Exploring Territorial and Temporal Patterns in News Coverage and Tone. *Journal of Common Market Studies*, 58(4):1034 – 1055, 2020.

[65] Judith Moeller, Damian Trilling, Natali Helberger, and Bram Es. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication and Society*, 21:1–19, March 2018.

[66] Philip M. Napoli. Deconstructing the Diversity Principle. *Journal of Communication*, 49(4):7–34, December 1999.

[67] OberCom. DIGITAL NEWS REPORT PORTUGAL 2023, 2023. Available at https://obercom.pt/wp-content/uploads/2023/06/DNRPT_2023_Final_15Junho.pdf, last accessed in October 2024.

[68] Kalyani Pakhale. Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges. arXiv preprint arXiv:2309.14084, September 2023.

[69] J. Peter. Agenda-Rich, Agenda-Poor: A Cross-National Comparative Investigation of Nominal and Thematic Public Agenda Diversity. *International Journal of Public Opinion Research*, 15(1):44–64, March 2003.

[70] Mark Raasveldt and Hannes Mühleisen. Duckdb: an embeddable analytical database. In *Proceedings of the 2019 International Conference on Management of Data (SIGMOD '19)*, page 4, Amsterdam, Netherlands, 2019. ACM.

[71] Danielle Raeijmaekers and Pieter Maeseele. Media, pluralism and democracy: what's in a name? *Media, Culture and Society*, 37(7):1042–1059, 2015.

[72] Shaina Raza and Chen Ding. Relevancy and Diversity in News Recommendations. In *CEUR Workshop Proceedings*, volume 3411, pages 6 – 15, 2022.

[73] Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4):1064–1082, October 2014.

[74] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM '15)*, pages 399–408, Shanghai, China, February 2015. ACM.

[75] Claude Shannon and Warren Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, 1949.

[76] Moses Shumow and Mercedes Vigon. News diversity and minority audiences: Using real simple syndication (RSS) to assess the democratic functions of Spanish-language media in the digital age. *Journalism Practice*, 10(1):52 – 70, 2016.

[77] E. H. Simpson. Measurement of diversity. *Nature*, 163:688, 1949.

[78] Helle Sjøvaag and Truls André Pedersen. The effect of direct press support on the diversity of news content in Norway. *Journal of Media Business Studies*, 15(4):300 – 316, 2018.

[79] Babak Sohrabi, Iman Raeesi Vanani, and Meysam Namavar. Investigation of trends and analysis of hidden new patterns in prominent news agencies of Iran using data mining and text mining algorithms. *Webology*, 16(1):114 – 137, 2019.

[80] Tobin South, Bridget Smart, Matthew Roughan, and Lewis Mitchell. Information flow estimation: A study of news on Twitter. *Online Social Networks and Media*, 31:100 – 231, September 2022.

[81] Fábio Souza, Rodrigo Frassetto Nogueira, and Roberto A. Lotufo. Bertimbau: Pretrained BERT models for brazilian portuguese. In Ricardo Cerri and Ronaldo C. Prati, editors, *Intelligent Systems - 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20-23, 2020, Proceedings, Part I*, volume 12319 of *Lecture Notes in Computer Science*, pages 403–417. Springer, 2020.

[82] Andy Stirling. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719, August 2007.

[83] Mikalai Tsytsarau, Themis Palpanas, and Malu Castellanos. Dynamics of news events and social media reaction. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 901 – 910, 2014.

[84] UB Mannheim. Reichsanzeiger-nlp: Ner/nel corpus for the german historical newspaper "deutscher reichsanzeiger und preußischer staatsanzeiger" (1819–1939), 2023. Available at https://ub-mannheim.github.io/reichsanzeiger-nlp/ last accessed in June 2025.

[85] Takeaki Uno, Hiroki Maegawa, Takanobu Nakahara, Yukinobu Hamuro, Ryo Yoshinaka, and Makoto Tatsuta. Micro-clustering: Finding small clusters in large diversity. *CoRR*, abs/1507.03067, 2015.

[86] Takeaki Uno, Hiroki Maegawa, Takanobu Nakahara, Yukinobu Hamuro, Ryo Yoshinaka, and Makoto Tatsuta. Micro-clustering by data polishing. In Jian-Yun Nie, Zoran Obradovic, Toyotaro Suzumura, Rumi Ghosh, Raghunath Nambiar, Chonggang Wang, Hui Zang, Ricardo Baeza-Yates, Xiaohua Hu, Jeremy Kepner, Alfredo Cuzzocrea, Jian Tang, and Masashi Toyoda, editors, *2017 IEEE International Conference on Big Data (IEEE BigData 2017), Boston, MA, USA, December 11-14, 2017*, pages 1012–1018. IEEE Computer Society, 2017.

[87] J. J. Van Cuilenburg, J. Kleinnijenhuis, and J. A. De Ridder. Towards a graph theory of journalistic texts. *European Journal of Communication*, 1(1):65–96, 1986.

[88] Jan Van Cuilenburg. On Competition, Access and Diversity in Media, Old and New: Some Remarks for Communications Policy in the Information Age. *New Media & Society*, 1(2):183–207, August 1999.

[89] Anita MJ van Hoof, Carina Jacobi, Nel Ruigrok, and Wouter van Atteveldt. Diverse politics, diverse news coverage? A longitudinal study of diversity in Dutch political news during two decades of election campaigns. *European Journal of Communication*, 29(6):668 – 686, 2014.

[90] Paul S. Voakes, Jack Kapfer, David Kurpius, and David Shano-Yeon Chern. Diversity in the News: A Conceptual and Methodological Framework. *Journalism & Mass Communication Quarterly*, 73(3):582–593, September 1996.

[91] Daniel Vogler, Linards Udris, and Mark Eisenegger. Measuring Media Content Concentration at a Large Scale Using Automated Text Comparisons. *Journalism Studies*, 21(11):1459–1478, August 2020.

[92] Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, and Daan Odijk. RADio* – An Introduction to Measuring Normative Diversity in News Recommendations. *ACM Trans. Recomm. Syst.*, 3(1), August 2024.

[93] Youzhong Wang, Daniel Zeng, Bin Zhu, Xiaolong Zheng, and Feiyue Wang. Patterns of news dissemination through online news media: A case study in China. *Information Systems Frontiers*, 16(4):557 – 570, 2014.

[94] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. In Lada A. Adamic, Ricardo Baeza-Yates, and Scott Counts, editors, *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. The AAAI Press, 2011.

[95] Nicholas J.Y. Wong and Hui Na Chua. A framework integrated with time series and text mining models to compare agenda-setting patterns of news and social media. *Journal of Engineering Science and Technology*, 17(3):1880 – 1896, 2022.

[96] Jianhua Yin and Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*, pages 233–242, New York, NY, USA, August 2014. ACM.

[97] Zhi-Qiang You, Yan-Yan Zhu, Xiao-Pu Han, and Linyuan Lü. Modelling temporal patterns of news report. In *Chinese Control Conference, CCC*, volume 2015-September, pages 1345 – 1350, 2015.

[98] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[99] Bo Zhang, Jinchuan Wang, and Lei Zhang. Exploring information flow patterns between news portals and microblogging platforms. In *Springer Proceedings in Complexity*, pages 187 – 199, 2013. Type: Book chapter.

[100] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web - WWW '05*, page 22, Chiba, Japan, 2005. ACM Press.