FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Towards an Automated Media Chart

Framing News Articles with Natural Language Processing Techniques

José Luís Sousa Tavares

Master in Informatics and Computing Engineering

Supervisor: Henrique Lopes Cardoso, University of Porto, PhD

Co-Supervisor: Rui Carreira, University of Maia, PhD

July 27, 2023

Towards an Automated Media Chart

José Luís Sousa Tavares

Master in Informatics and Computing Engineering

July 27, 2023

Abstract

News articles are the bridge between individuals and information. People use the news to inform themselves about contemporary topics and political cases. Recently, prominent events have materialized infodemic tendencies, with massive amounts of information becoming quickly open to the public. News coverage may severely impact public perception of events and affect opinions and decision-making. The varied news sources expose readers to different narratives and discordant judgments of the same fact. Media bias poses a severe problem to society because it requires readers to be adequately aware of the often minute though effective slants in the news or whenever coverage is intentionally distorted to influence general belief.

It is vital to entrust newsreaders to evaluate the news critically to face the problems arising from media bias. Manual approaches and techniques are highly effective. However, manual analysis often comes with enormous effort and demands consumers to have significant knowledge of media literacy practices and political science. Consequently, applying this manual effort to daily news consumption denotes an unbearable barrier.

This dissertation proposes an interdisciplinary approach to automatically detect and classify media slants and metrics in Portuguese news articles. Computerized methods enable a briefer and more reliable analysis of the potential biases and unreliability of the news. A better understanding of the news will stimulate better addressing of the public interest during a forthcoming crisis of news coverage.

This work gives readers a framing of news articles with four metrics – reliability, political stance, objectivity, and readability. The process of classifying each one is analyzed and described. The approaches employed use Natural Language Processing techniques, taking advantage of pre-trained BERT models or well-known formulas. A prototype is designed, described, and implemented to visualize the framing results. This results in a novel solution to visualize media bias dimensions in a browser-based platform that allows readers to visualize and interact with the framing information, allowing them to understand the disposition of the news in the Portuguese panorama.

Resumo

As notícias são a ponte entre os indivíduos a informação. As pessoas usam as notícias para se informar de tópicos correntes e acontecimentos políticos. Recentemente, eventos globais materializaram tendências infodémicas, com uma enorme quantidade de informação a ficar disponível aos consumidores. As notícias podem afetar severamente a perceção dos acontecimentos por parte dos consumidores e a opinião e decisão pública. As várias fontes de notícias expõem os leitores a diferentes narrativas e a informação discordante do mesmo evento. O viés na média coloca-se como um grave problema para a sociedade pois requer que os leitores estejam devidamente informados dos comuns vieses minuciosos, mas decisivos ou de quando a informação é intencionalmente alterada para alterar a convicção do público.

É necessário incentivar os leitores a avaliar as notícias de uma forma critica de modo a combater os problemas decorrentes do viés na média. Abordagens manuais para resolver esse problema são altamente eficazes. No entanto, as mesmas abordagens manuais exigem um esforço enorme e exige que os consumidores tenham conhecimento de práticas de literacia da média e ciência política. Consequentemente, aplicar esse esforço manual ao consumo de notícias indica uma barreira difícil de superar.

Esta dissertação propõe uma abordagem interdisciplinar para detetar e classificar automaticamente as inclinações de notícias portuguesas. Os métodos computorizados permitem uma análise mais breve e confiável dos possíveis vieses e falta de confiabilidade das notícias. Uma melhor compreensão das notícias estimulará uma melhor interpretação do público durante uma próxima crise de cobertura jornalística.

Este trabalho oferece aos leitores um enquadramento de notícias em quatro métricas - confiabilidade, posição política, objetividade, e legibilidade. O processo de classificação é analisado e descrito. As abordagens usadas recorrem a técnicas de processamento de linguagem natural, tomando vantagem de modelos pré-treinados BERT e fórmulas consolidadas na literatura. Um protótipo é desenhado, descrito, e implementado para visualizar os resultados do enquadramento das métricas. O resultado é uma solução nova para visualizar dimensões de viés nas notícias numa plataforma *browser-based* que permite que os leitores interajam com a informação, permitindo que os leitores compreendam a disposição das notícias no panorama Português.

“With enough of us, around the world, we’ll not just send a strong message opposing the privatization of knowledge – we’ll make it a thing of the past”

Aaron Hillel Swartz

Contents

1 Introduction **1**
1.1 Problem . . . 1
1.2 Research Gap and Motivation . . . 3
1.3 Research Question and Objectives . . . 4
1.4 Dissertation Structure . . . 5

2 Media Bias **6**
2.1 Media Bias . . . 6
2.1.1 Definitions . . . 7
2.1.2 Forms of Media Bias . . . 8
2.1.3 The Societal Effect . . . 8
2.2 Media Bias by Word Choice and Labeling . . . 10
2.3 Approaches to Identify Media Bias by Word Choice and Labeling . . . 11
2.3.1 Manual Identification . . . 11
2.3.2 Automated Identification . . . 12
2.4 Related Work on Framing Platforms . . . 14
2.5 Key Findings . . . 15
2.6 Summary of the Chapter . . . 15

3 News Articles Framing Approach **17**
3.1 Introduction . . . 17
3.2 Identification of Axes of Interest . . . 18
3.2.1 Reliability . . . 19
3.2.2 Political Stance . . . 19
3.2.3 Objectivity . . . 20
3.2.4 Readability . . . 21
3.3 Framing Approach . . . 21
3.3.1 Datasets . . . 22
3.3.2 Data Collection . . . 23
3.3.3 Task . . . 27
3.4 Prototype . . . 30
3.5 Summary of the Chapter . . . 34

4 Axes-oriented Automatic Classification **36**
4.1 Introduction . . . 37
4.2 Reliability . . . 37
4.2.1 Related Work . . . 37
4.2.2 Datasets . . . 38

4.2.3 Exploration 39
4.2.4 Method 42
4.2.5 Evaluation 45
4.3 Political Stance 46
4.3.1 Related Work 47
4.3.2 Datasets 48
4.3.3 Exploration 50
4.3.4 Method 51
4.3.5 Evaluation 53
4.4 Objectivity 55
4.4.1 Related Work 55
4.4.2 Datasets 55
4.4.3 Method 56
4.4.4 Evaluation 56
4.5 Readability 58
4.5.1 Readability Indices and Formulas 58
4.6 Future Work 62
4.7 Summary of the Chapter 63

5 Prototype 65
5.1 Introduction 65
5.2 Use Cases 66
5.2.1 List of News Articles, the First Interaction 66
5.2.2 News Article Page, the Detailed Interaction 67
5.2.3 Chart, the Interactive Interaction 68
5.2.4 List Collections and Models, the Informative Interaction 71
5.3 Architecture 71
5.3.1 Backend 72
5.3.2 User Interface 75
5.4 Future Work 75
5.5 Summary of the Chapter 76

6 Conclusions 77
6.1 Summary 77
6.2 Threats to Validity 78
6.3 Contribution 79
6.4 Future Work 80

References 82

A Data Collection Files 89
A.1 Brexit Configuration File 89
A.2 Opinion Articles Configuration File 89
A.3 Reliable Articles Configuration File 90
A.4 Unreliable Articles Configuration File 91
A.5 COVID-19 Articles Configuration File 93
A.6 Example Output File 95

B Arquivo.pt API Requests Examples 96

B.1 Arquivo.pt CDX Server API Request 96
B.2 Arquivo.pt Full-text search API Request 97

C OpenAI Chat Completion API 100

List of Figures

2.1 News creation process from an event to the consumer’s perception, adapted from [1]. 7

3.1 Flowchart of the data collection process. 24
3.2 Ad Fontes Media interactive chart [2]. 31
3.3 AllSides headline roundup with news about the missing Titanic sub. 32
3.4 The Factual news coverage of the Supreme Court rejection against Navajo Nation. 33

4.1 Average number of organization, location, and person entities per article in the training dataset. 40
4.2 Proportion of sentiments per label in the training dataset. 42
4.3 Different approaches, (1) and (2), to feature concatenation 45
4.4 Proportion of sample’ topics per label in the political stance dataset using LDA with tf-idf. 50

5.1 Main page of the web prototype. 68
5.2 Individual page of a news article in the web prototype. 69
5.3 Individual page of a news article in the web prototype. The framing detailed are expanded in the green square box. 70
5.4 Radar chart example with two media outlets in the web prototype. 72
5.7 Architecture of the prototype. 72
5.5 Bi-dimensional chart example in the web prototype. 73
5.6 Collections and models page in the web prototype. 75

List of Tables

1.1 Two news sources reported the same event on March 11, 2003. 2

2.1 Overview of the bias forms adapted from the identification made by Park et al. [3] and Hamborg [1]. 9

3.1 Gathered datasets for the classification of the media axes. 22
3.2 Examples of Brazilian Portuguese, European Portuguese, and multilingual transformer-based pre-trained models. 29

4.1 Datasets considered for the reliability classification task. 39
4.2 LDA and BERTopic topic modeling for the reliability dataset. The results show the five topics with the most documents. 41
4.3 Reliability models' configuration using method 1. 43
4.4 Extra features, their description, and type, used in the method feature concatenation. 44
4.5 Architectures to fine-tune the models using feature concatenation. 46
4.6 Results obtained for the reliability classification using BERTimbau, XLM-RoBERTa, and Albertina-ptpt. 46
4.7 Portuguese political parties and their respective strict and relaxed labels derived from Wikipedia's party label. 49
4.8 Strict label and relaxed label distributions in the political dataset. 49
4.9 LDA and BERTopic topic modeling for the political stance dataset. The results show the five topics with the most documents. 51
4.10 Political models' configuration. 52
4.11 Absolute distances between the labels on the political dataset. 53
4.12 Results obtained for the political stance classification with the strict labels using BERTimbau, XLM-RoBERTa, and Albertina-ptpt. 54
4.13 Results obtained for the political stance classification with the relaxed labels using BERTimbau, XLM-RoBERTa, and Albertina-ptpt. 54
4.14 Label and relaxed label distributions in the objectivity dataset. 55
4.15 Absolute distances between the labels on the objectivity dataset. 56
4.16 Objectivity models' configuration. 57
4.17 Results obtained for the objectivity classification using BERTimbau, XLM-RoBERTa, and Albertina-ptpt. 57

5.1 Backend's REST API endpoints, their description, and whether they have an implemented UI in the prototype. 74

List of Algorithms

1 Algorithm to estimate the number of syllables in Portuguese words. 60

2 Algorithm to convert a list of probabilities into a continuous value between -1 and 1. 71

Listings

A.1 Example of a configuration file to search the terms “Brexit” and “European Union” in the The Guardian website between January 1st 2019 to December 31st 2020. . 89
A.2 Example of a configuration file to search the terms related to “opinião” in a few Portuguese news websites between January 1st 2018 to December 31st 2019. . . 89
A.3 Example of a configuration file to search diverse topics in a few Portuguese reliable news websites between January 1st 2010 to December 31st 2022. 90
A.4 Example of a configuration file to search diverse topics in a few Portuguese fake and unreliable news websites between January 1st 2010 to December 31st 2022. . 91
A.5 Example of a configuration file to search the terms related to COVID-19 in a few Portuguese news websites between January 1st 2020 to December 31st 2022. . . 93
A.6 Example of an output file of the data collection tool. 95
B.1 Arquivo.pt CDX Server API curl command to get the preserved pages as a JSON from sapo.pt in January 1st, 2023. 96
B.2 Query JSON response to get the preserved pages as a JSON from sapo.pt in January 1st, 2023. 96
B.3 Arquivo.pt Full-text search API curl command to get one article matching “tribunal” from “publico.pt”. 97
B.4 Arquivo.pt Full-text search API response to get one article matching “tribunal” from “publico.pt”. 97
C.1 OpenAI’s Chat Completion API request example using data from the political stance dataset. 100
C.2 OpenAI’s Chat Completion API reply to the example above. 101

Abbreviations

ARI	Automated Readability Index
BERT	Bidirecional Encoder Representations from Transformers
NER	Named Entity Recognition
NLP	Natural Language Processing
WHO	World Health Organization

Chapter 1

Introduction

1.1 Problem 1
1.2 Research Gap and Motivation 3
1.3 Research Question and Objectives 4
1.4 Dissertation Structure 5

This chapter serves to present and clarify the problem addressed in this work, as well as provide motivation for the research. To begin, Section 1.1 highlights the issue of biased news coverage through a concise introduction supplemented by examples and foundational information. By doing so, it establishes the importance and relevance of the topic. Subsequently, Section 1.2 identifies the existing gap in the current literature and establishes the motivation behind this work. It elaborates on why further investigation is needed and how this research aims to fill that gap. Moving forward, Section 1.3 defines the specific research questions and objectives of the dissertation. These research questions will serve as guiding directions for the entire study and the subsequent development of the work. Lastly, in Section 1.4, the document's structure is outlined, providing a roadmap of what to expect in the subsequent chapters.

1.1 Problem

The news is the foremost source of information on events and contemporary topics. Millions of citizens daily turn to the news media to inform themselves. Therefore, media plays a crucial role in society. How an event or political case is portrayed in the news influences how people perceive events and might alter their decision-making or opinion in public debates [1]. Often, news consumers are exposed to multiple narratives of the same event and occasionally to contradicting ones [4]. The principal source of information for millions of people must become reliable and neutral. However, that is often not the case, and news sources frequently express biases, commonly called *media bias*, that distort information or influence general beliefs and judgments.

Biases can occur in multiple forms, more explicitly and easily detected or more meticulously and challenging to identify. Selective omission of facts, choice of words, limiting debate, and story framing are effective techniques in this matter [5]. Nowadays, it isn't easy to find a completely unbiased news source. News sources tend to accommodate the information they provide to go in the direction of consumers' prior beliefs. Or, for example, suppose that scientists have devised a way to work out cold fusion. News sources might hesitate to report such discovery because consumers' prior beliefs make them believe that cold fusion is unlikely. Thus, consumers will infer that the news sources may report poor information [6]. With the same rationale, news sources make room for intentional biases to maintain the already-fitted readers.

As seen, news sources can induce intentional biases to accommodate readers. However, and perhaps the most alarming component of biases, the slants are deliberately introduced to change consumers' minds about a particular topic in particular circumstances. Multiple studies have found that some biases are harmful and might affect worrying issues such as democratic elections [7, 8]. In 2006 a study showed that American news outlets had a solid liberal bias [8]. In 2007 a survey of the 2000 American elections concluded that Fox News had increased 0.4 to 0.7 the percentage of the Republican vote share [7].

Table 1.1: Two news sources reported the same event on March 11, 2003.

Source	**Passage**
The New York Times	Iraqi fighter jets threatened two American U-2 surveillance planes, forcing them to return to abort their mission
USA Today	U.N. arms inspectors said Tuesday they had withdrawn two U-2 reconnaissance planes over Iraq for safety reasons

In Table 1.1, it is observed the way two distinct news sources reported the same incident on March 11, 2003. With different word choices, The New York Times makes the readers believe that Iraqi fighter jets forced the surveillance planes to abort the mission. On the other hand, USA Today passed the message that the aircraft had to be withdrawn for safety reasons without mentioning Iraqi fighter jets.

With the fear that news sources become more and more relevant to such decisive matters, it is essential to be aware of such biases. Today, biases can materialize in slanted news and partially fake or utterly fake news, making what consumers catch sight of less and less reliable [9, 10]. It is not easy or trivial for the regular news consumer to recognize such cases. Identifying such issues requires appropriate means and knowledge of media literacy. The technological era brought new notions of literacy. Nowadays, the domains of literacy span multiple fields, such as computers, multimedia, and media. At the beginning of this century, anticipating the growing interest in media literacy, some definitions of it and its core concepts were developed [11]. Yet, education and society seemed not to adapt appropriately to the newer needs as they are broader, resulting in a lack of media literacy when arguably it is more critical than ever [12].

The problem of slanted news gets worse when the number of news increases. It is a linear problem – the more information, the more significant number of slanted ones. Global events tend to create new infodemic tendencies. Recently, with the outbreak of the COVID-19 pandemic, the world had a never seen amount of news produced every day. Health organizations' concern was to stop the outbreak and fight the infodemic tendency. In 2020, the WHO (World Health Organization) director said, "'We're not just fighting an epidemic; we're fighting an infodemic'...". On a smaller scale, elections, sporting events, or political crises can also create dangerous infodemic tendencies [13].

1.2 Research Gap and Motivation

News consumption has become an inherent part of people's daily routines. However, the quality and integrity of news have been called into question, with media having a significant say in shaping public opinion. Prominent global events may materialize infodemics. Excessive loads of information will widen the gap between a consumer and an utterly trustworthy news source. That said, the motivation of this research is to entitle consumers with a way to assess the quality and attributes of the news.

Manually evaluating the quality of the news is reliable, yet a very resource-intensive task due to the extent of news available [1]. In the technological era, computers put forward an opportunity to tackle the issue. In computer science, media bias is a relatively young research topic with gaps to be covered [14, 1] and there is an ongoing movement toward reliably identifying these biases. With the advances in natural language processing (NLP), it is now possible to identify specific slants veraciously.

Multiple companies have dedicated their efforts to addressing media bias, particularly in the United States. AllSides[1] and The Factual[2] are notable examples of companies that strive to detect and educate news consumers about media biases. However, the research conducted in the Portuguese context is significantly limited or virtually non-existent. While fact-checking initiatives have emerged within some Portuguese media outlets to inform the public about unreliable news coverage, this represents the closest area of study that can be found. The scarcity of research in this domain presents opportunities for further development and investigation.

Therefore, the motivation is to fill the gap and address the problem of media bias and the lack of information available to news consumers about the quality and attributes of news sources. Through an automated approach, news consumers will be empowered with the necessary tools to make informed decisions and have selective behavior when consuming news.

[1] allsides.com

[2] thefactual.com

1.3 Research Question and Objectives

This research will focus on the Portuguese media panorama to narrow the media bias problem, particularly on news sources that report events with textual news. Having to acknowledge the research above gap and motivation, the research questions of this dissertation are defined as:

Question 1 How can natural language processing techniques help news consumers understand media and selectively consume news?

Question 2 Which media slants and attributes help news consumers understand media?

In the context of question 1, natural language processing enables computers to handle text data, such as the text in news content. This computer science field can treat news data and perform several tasks useful in slant framing. With NLP techniques, it is possible to identify specific slants or attributes that news exposes and report them to news consumers. The previous leads to the second question: the slants and features that help consumers understand the information. For this intention, this work proposes four axes, which are slants or attributes of news, which can help consumers understand the direction of news articles. The axes are reliability, political stance, objectivity, and readability. They were identified after research on the nature of news, realized in Chapter 2 and are later detailed in Chapter 3.

The four axes help consumers understand the panorama of the news sources by assessing which orientation of these attributes of the news they want to focus their reading. The framing of the information aims to yield a platform where consumers can overview and compare the framing results from one news source to another.

Having answered the previous research questions, it is possible to define the next one:

Question 3 How can these axes be presented to news consumers?

To address this question, it is essential to research the available resources on the internet. As mentioned previously, AllSides and The Factual are well-known companies that actively tackle the issue of media bias. These platforms curate news articles on their websites and give readers insights into how they have classified the content. In Chapter 2, an investigation is conducted on these platforms, as well as other media bias platforms, to gather valuable information on how news articles can be effectively framed for news consumers. These platforms present similar interfaces and offer streamlined approaches to present information while swiftly ensuring an exceptional user experience. Consequently, by analyzing the foundational work carried out on these platforms, the answer to this question will come into focus.

All in all, the objective is to use NLP techniques to frame Portuguese news with pre-defined axes, which will help consumers understand the scenery of the news sources and qualitatively assess their sources of information.

1.4 Dissertation Structure

Chapter 2 performs a literature review on media bias. It outlines the multiple media bias definitions, forms, and societal effects. It also investigates approaches to detecting media bias and platforms that inform news readers about it. Then, in Chapter 3, the study defines the problem solution by pointing out the four axes of interest to understand news articles. It also describes the prototype to visualize the framing of each axis. Then, Chapter 4 explains the methods used to frame each axis. The methods comprise state-of-the-art BERT models for text classification and useful text metrics for readability estimation. Chapter 5 details the prototype, its use cases, and its architecture. As a prototype, it suggests further improvement to create a working, ready-to-deploy web platform for media bias news coverage. Lastly, Chapter 6 provides the investigation findings, outlining the main contributions and explorations. It also presents the threats to the validity of this research and future work.

Chapter 2

Media Bias

2.1 Media Bias 6
2.1.1 Definitions 7
2.1.2 Forms of Media Bias 8
2.1.3 The Societal Effect 8
2.2 Media Bias by Word Choice and Labeling 10
2.3 Approaches to Identify Media Bias by Word Choice and Labeling 11
2.3.1 Manual Identification 11
2.3.2 Automated Identification 12
2.4 Related Work on Framing Platforms 14
2.5 Key Findings 15
2.6 Summary of the Chapter 15

This chapter provides an interdisciplinary literature review on media bias. It gives an introduction to media bias and its scrutiny. After, it shows the definition of bias and how it has a powerful influence on society. Then, it aggregates the approaches to identify biases, proceeding from early procedures to state-of-the-art ones. Next, the chapter shows research on current solutions to identify the biases. Finally, it ends with the key findings of the literature review and how they will strive for the problem solution.

2.1 Media Bias

Media bias is not a recent research topic. The first wave of analysis and exploration of the matter came in the second half of the 20th century. Most of the research done had the motivation of bias influence in democratic elections [15, 16, 7, 13]. Alongside the analysis, multiple definitions of media bias were defined by many researchers. Today, there is no consensual definition of media bias in the social sciences, but different meanings depending on the research topic. Many studies on the influence of biases in society have been conducted as they showed an alarming issue to the public.

The news production process is extensive and passes through many sectors and persons, from event gathering and selection to actual writing. It is essential to understand the process where media bias might arise. The first step of the process is where the events and facts are gathered. The bias can upturn in this part of the process, primarily due to fact and source selection. Depending on the piece, information can be omitted. The next step is the actual writing of the data. This step is conditioned by word choice and labeling, possibly leading to biases. This research will focus on this step of the news creation process. The last step before reaching consumers is the presentation step. In this step, the editor is challenged with issues such as picture selection and explanation, size allocation, and placement [17].

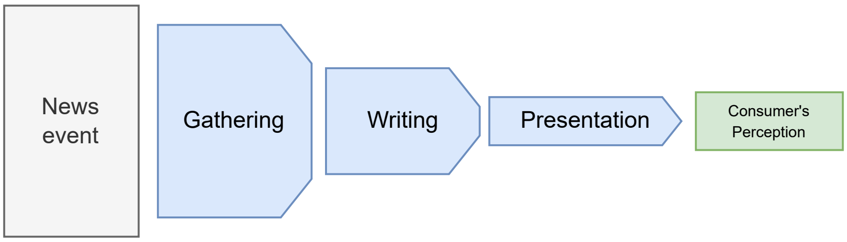


Figure 2.1: News creation process from an event to the consumer's perception, adapted from [1].

The biases surge in the gathering, writing, and presentation phases in the news creation process. However, the consumer's perception is also responsible for augmenting or diminishing their intensity. It depends on multiple factors, including background knowledge, social status, and country [1].

2.1.1 Definitions

Since the start of the research on media bias, researchers have proposed many definitions. Some definitions are more narrow than others and depend on the purpose of the study in which they were defined. In 1975, Williams [18] defined bias as volitional, putatively influential, reasonably, and sustained. Bias has to be willful, have influence, and be something that lasts and is not one-shot. Occurrences where the acts are insufficient to be influential or are not plausible, *e.g.,* threats of conventional values or extreme biases (*e.g.*, clearly identified gender bias), should not be considered bias. Later, D'Alessio et al. [15] studied the nature of media bias and defined media bias in three components: gatekeeping, coverage, and statement. Gatekeeping is the issue of selecting from a set of potential stories the ones that will be reported and those that will not. Coverage bias is concerned with the amount of coverage that topics receive. For example, in physical newspapers, coverage bias issues the happening of the aggregate information available for a specific topic. Usually, it is measured in column inches. Statement bias relates to presenting and writing the issues and focuses on whether the media coverage is "favorable" or "unfavorable" reporting a situation. Despite the research pointing out that these biases are hard to measure outside the electoral prospect, the three-component topology is well established in the media bias debate and motivated other studies [16, 1].

To narrow the issue of media bias, the definition used in this research is illustrated as the characteristics of news articles that help consumers understand their anatomy. The characteristics are specific dimensions of bias in the news writing style and content, such as reliability or political stance. The literature review of media bias supported identifying the most significant and helpful features of media bias. Overall, the characteristics identified, also named axes in this study, are political stance [13, 19, 16, 18], objectivity [16, 20, 21], and reliability [22, 9, 23, 24]. In Chapter 3, these axes are more in-depth detailed.

2.1.2 Forms of Media Bias

As seen in Figure 2.1, the news creation process has multiple stages. In the various phases of the process, diverse forms of bias surge. This section uses an adapted version of Hamborg's [1] bias forms identification to briefly describe the multiple forms of media bias. The bias forms identified are part of the news creation process defined by Park et al. [3] and later adapted by Hamborg [1]. Since this research aims to focus on the textual elements of news articles, the study does not further develop the forms of bias that arise in non-textual news articles or outside the writing phase of news articles.

Table 2.1 briefly describes the many forms of bias. The "Medium" column describes where the bias is present, and the "Target Object" identifies which target concept the bias affects.

2.1.3 The Societal Effect

Media bias has broad-ranging consequences today, affecting how individuals perceive and comprehend events. Media bias can further worsen the case in moments of crisis, political or electoral events, leading to misinformation and confusion. The significance of accurate and unbiased news reporting has never been more apparent, given how influential and far-reaching the effects of media bias can be on society. Media bias's influence on public opinion and events evolution can be observed using examples of crises, political, or electoral scenarios.

In the premature steps of the 1990s decade, the Persian Gulf war arose between Middle Eastern countries, notably Iraq and Kuwait. The political and military tension was the consummation of economic, political, and territorial disputes that led to the two-year war between 1990 and 1991. The fight was heavily televised and had significant media coverage, with several people reporting "staying up late" to watch the clash coverage [25]. According to the findings of the studies, the Gulf crisis news coverage carried Americans new ways to assess George Bush. Before the crisis, Americans were most apprehensive about economic and criminality concerns. Following the Iraqi invasion, the evaluations of the former US president also became dependent on foreign policy considerations [25]. Other studies noted a curiosity in getting politically knowledgeable

Table 2.1: Overview of the bias forms adapted from the identification made by Park et al. [3] and Hamborg [1].

Name	Medium	Target Object	Phase	Explanation/Example
Event Selection	News outlet	News article	Gathering	News outlets systematically ignore negative events regarding a topic (e.g., not reporting news criticizing the government).
Source Selection	News article, text	Text, picture	Gathering	Voluntarily selection of multiple sources that report a specific perspective of an event.
Commission and Omission	News article, text	Text	Gathering	Facts that support a perspective are added, or facts that undermine the perspective are omitted.
Word Choice and Labeling	Text	Entity, action, attribute, etc.	Writing	Selective word choice to label a specific concept. “illegal alien” as “undocumented migrant”. Negative words in the vicinity of a concept (e.g., Morroco fans as “Morroco rioters” or “Wild Morroco fans”).
Story Placement	News outlet	News article	Editing	Front page story covers more attention than column story.
Size Allocation	News outlet	News article	Editing	A large story receives more attention than a short story.
Picture Selection	News article, picture	Picture, entity, action, etc.	Editing	Selection of pictures that represent fleeting moments to distort reality.
Picture Explanation	Text	Picture caption	Editing	Caption puts the picture into context or distorts it.
Spin	News article, outlet	One or more news articles	All phases	The overall slant of the news article. It is the result of multiple biases combined.

in wartime. On the other hand, regular news consumers cannot do so if they have no preceding political background [26], which may lead to mistaken premises.

Other studies have been conducted in media coverage fields, such as electoral cases. In the decade of 2000s, numerous studies on the American electoral process surfaced [7, 15, 8]. In 1996 the twenty-four-hour Fox News Channel was introduced. Fox News expanded, and in June 2000, they had an audience of 17.3 percent [7]. DellaVigna et al. compared the Republican vote share between 1996 and 2000 in the towns that had adopted Fox News. The study concluded that the rise of Fox News in some news resulted in an increase of 0.4 to 0.7 percent of the Republican vote share in the 2000 elections. Early studies of presidential elections concluded that there was no significative partisan bias, *i.e.,* bias favoring either Republicans or Democrats. This does not imply that the news reports were fair. Conversely, there is commonly no discernible net bias, yet they do exist, but they also happen to be punctual and balance one another [15].

2.2 Media Bias by Word Choice and Labeling

This research is directed toward understanding media with the textual content of news. Many forms of media bias exist, as described in Table 2.1. However, this research aims to identify media bias that arises in the writing process of news articles, particularly the word choice and labeling form of bias. The word choice and labeling form of bias is thus introduced in this section.

Authors select words to label a *concept* in a news article's writing process. Authors may be unaware of bias-inducing terms in this process, yet in some cases, they are carefully selected to induce bias. A semantic concept may be an entity, geographical position, event, or even a more abstract concept, such as a software program. Instances of word choice and labeling bias frame the identical concept differently in distinct contexts. Depending on the context, this can have a positive or negative connotation for the concepts. Common examples used in this analysis include "immigrant" or "economic migrant" [1] and "illegal alien" or "undocumented immigrant" [27].

Authors usually label concepts using conventional category labels. People continuously confirm and maintain the meaningful categorization of concepts by systematically using the same labeling. For example, it is common to label age groups (*e.g.*, elderly, adolescent) or nationalities (*e.g.*, Portuguese, Italian). However, over time, arbitrary characteristics could aggregate people [28]. At the beginning of the internet, there was no term for people working remotely and traveling the world, working from many locations, and having the freedom to do so. However, when a concept, in this case, a group of people, becomes a topic of discussion, people will likely find terms to label such a concept unconventionally. The example of people working remotely from many locations was later tagged "digital nomad", which denotes a person that uses technology, primarily the internet, to earn a living working remotely with the freedom to work and live anywhere in the world.

Unconventional concept labeling is not necessarily wrong, as the need to label such topics appear. However, once a concept is linguistically labeled, it achieves reality as it will more likely gain value. When concepts, particularly new concepts, gain significance, news outlets tend to construct a judgment or a stereotype around it [28]. The nature of the judgment yielded in the concept is often characterized by the words in the vicinity of it (*e.g.*, adjectives) [29]. Following Morroco's World Cup 2022 match win against Belgium, riots emerged in Brussels. The Brussels mayor Philippe Close said, "Those are not fans, they are rioters (...)" [30]. Few news outlets generalized Morrocan fans as rioters and savages, clearly stereotyping them using slanted coverage.

2.3 Approaches to Identify Media Bias by Word Choice and Labeling

In an age where information is widely accessible but often unreliable, manual and automated approaches to identifying media bias have become highly important. With the rise of fake news and the potentially harmful effects of media bias, such as the ones described in Section 2.1.3, it is essential to develop methods to distinguish between reliable and slanted news sources. Researchers have developed manual and, more recently, automated ways to identify media bias in news articles.

2.3.1 Manual Identification

The manual approaches to identifying media bias have roots in the social sciences. Researchers have conducted content analyses to identify and quantify media bias in coverage. The research typically focuses on two studies content-oriented and frame oriented. This subsection describes both analysis types and approaches used to identify media bias.

The *content analysis* identifies media bias by detecting and quantifying its instances in news texts. The study defines the questions and hypotheses that might surge in a pre-gathered corpus of data. Then, annotators, also named coders, read the news articles and decide whether to accept or reject the initial questions and hypotheses. There are two distinct methods for content analysis: deductive content analysis and inductive content analysis.

Both of the methods either rely on a codebook or devise a codebook. A codebook is a kind of instruction manual for bias annotation. It contains definitions, rules, examples, and directives on how and what to annotate. In a deductive analysis, annotators have an a priori codebook. Then annotators methodically read the news in the corpus and annotate the information. In an inductive study, the coders read the corpus without having instructions or directives, only knowing the research questions and hypotheses. This type of research is used to verify investigation and patterns or to devise a codebook [1].

Frame analysis investigates how media sources portray information and how that presentation impacts public perception. This method examines how the media frames issues, events, public opinion, and attitudes. Researchers may explore characteristics such as the reporting tone, language, and overall narrative. The research on frame analysis does not converge in only textual information in news articles. It transcends multiple forms of media bias. The example of Fox

News' impact in the presential elections [7] in Section 2.1.3 is an example of how framing analysis can be used on other media bias forms.

Both analyses can accomplish the detection of media bias. Thus, using both methodologies will improve the analysis's results. Performing analysis and combining the results is often called *meta analysis*.

Identifying bias using content analysis is the most common approach in the social sciences and delivers promising results. However, the manual calculation is time-consuming and requires excellent expertise [1]. For instance, manual methods have revealed that the New York Times uses a more dramatic tone than The Washington Post [31].

2.3.2 Automated Identification

While social science researchers turn in the direction of manual approaches, computer scientists try to solve the excessive effort caused by manual procedures with automated solutions. This section provides a literature review of general practices to identify media bias by word choice and labeling in news sources. The research focuses on methods for detecting media bias and identification methods. Detection methods identify whether a sentence or a set of sentences contains biased information and possibly quantify it on a scale. On the other hand, identification methods outline which parts of a text sequence are biased.

The approaches elucidated in this section contain general methodologies for detecting and discerning media biases. Each study presents its distinct definition of media bias. These approaches follow a similar trajectory to the one described in this work, which involves scrutinizing specific attributes of news articles and studying them to identify bias. While this research essays a comprehensive examination of the problem by dividing it into four distinct axes, the works described below typically approach media bias as a singular dimension. They do not investigate the finer components of media bias to understand them better. Nevertheless, these general approaches offer valuable insights into detection mechanisms that will prove advantageous in the subsequent stages of this research.

As online news and information volume grow, automated media bias detection approaches have become increasingly popular. In computer science, natural language processing is concerned explicitly with understanding human text and speech. Hence, words, phrases, or pieces of content of news articles can be further analyzed using NLP techniques to detect media bias by word choice and labeling. In that matter, the sentiment analysis sub-field of NLP is the closest field. Sentiment analysis is the topic that detects emotional tones in a textual sequence. Additionally, other features apart from sentiment can be extracted by analyzing the text's tone. Still, sentiment analysis has shown liabilities, namely (1) lack of gold-standard datasets and (2) the sentiment-inducing phrases are likely to have high context-dependency.

Multiple studies have explored the use of automated approaches. In 2011 a survey conducted by Diakopoulos and Shamma [32] showed that in the 2008 presidential elections, the tone of the media coverage was significantly more positive towards a candidate. The study performed a sentiment analysis on the tweets posted during the debates. Similarly, Potthast et al. [33] analyzed a large corpus of news articles and found characteristic pieces of evidence that distinguish articles conveying a hyperpartisan view. The study also evidences that left and right extremism often use the same writing style.

The approaches to detect biases have been in constant evolution. Lately, there has been a change in the direction of the algorithms used in the detection from machine learning algorithms to deep learning algorithms. The approaches used traditional machine learning algorithms with feature engineering in prior studies. Spinde et al. [34] tested various machine learning algorithms with state-of-the-art feature engineering techniques, the most notable being a semi-automated bias lexicon introduced by Hube and Fetahu in 2018 [35]. The approach employed correctly identified around 77% of biased words and was used to identify media bias in news articles.

Transformers [36] in deep learning changed the paradigm by introducing a new architecture that could process sequential data effectively and efficiently. This innovation allowed Transformers to achieve state-of-the-art results in various NLP tasks. A few times later, Google researchers introduced BERT (Bidirectional Encoder Representations from Transformers) [37]. BERT is a Transformer-based language model with state-of-the-art results in diverse NLP tasks. Since the surge of Transformer-based models, most NLP tasks solved by traditional algorithms became outperformed and deprecated. That brought a research gap in the fields where the conventional models were employed and the enthusiasm to replace deprecated models with Transformer-based ones [38].

With the advancements made with the BERT models, many researchers have proposed works that aim to detect and solve the media bias identification problem with the Transformer-based model [39, 40, 38, 41, 42]. Particularly, the work done by Spinde et al. in [40, 38, 42] represents greats advancements in the research by introducing two gold-standard datasets annotated by experts - MBIC [42], and BABE [40]. Krieger et al. [41] developed a BERT-based model to detect bias by word choice over the BABE dataset. The model can identify sentence-level bias state-of-the-art scores.

BERT models have demonstrated exceptional performance in various tasks, prompting researchers to explore numerous avenues to further enhance or expand these models for even better results [43, 44]. Among these approaches, one particularly beneficial technique for text classification tasks involves studying the data to determine the feasibility of extracting additional information. By combining engineered features with BERT, a rich method emerges that capitalizes on the outstanding performance of BERT while incorporating valuable information that would otherwise be inaccessible without preliminary analysis.

In short, while social science researchers prefer manual approaches, computer scientists focus on automated solutions. The research examines detection and identification methods, and although this work takes a comprehensive approach, the described approaches offer valuable insights. Automated approaches using natural language processing (NLP) techniques like sentiment analysis have gained popularity in detecting media bias. Transformer-based models like BERT have revolutionized the field, outperforming traditional methods. Researchers have proposed BERT-based models to address media bias, achieving advancements in identifying biased content.

2.4 Related Work on Framing Platforms

This section helps to position the current research within the context of the body of knowledge and point out the gaps and limitations of the previous studies. Apart from the earlier sections, this section provides examples of related work in media bias detection, whether manual or automated or inside or outside the scope of word choice and labeling.

NewsCube [3], and later NewsCube2.0 [45] are social news websites created to mitigate media bias. Users collectively create a framing spectrum on the website to give a global view of the news producers. The focus of NewsCube2.0 is the crowdsourcing system for framing in the news and ways to visualize the framing. The system, however, lacks automation as it is directed to crowdsourcing annotation, thus representing a significant drawback in bias detection.

Another example of previous work in the media bias area is fact-checkers. Fact-checkers position themselves in assessing whether media coverage is reported factually or not. They are primarily concerned with a specific type of slant: the reliability of facts. The approaches fact-checkers use are manual and require efforts outside the form of biased word choice and labeling. Polígrafo [46], and Observador Fact Check [47] are two Portuguese fact checkers associated with news outlets, SIC and Observador, respectively. Essentially, they look for spin in popular news or information that can be biased to check the factuality of the data. That is used to help the Portuguese population understand what information they can rely on.

In the media domain in the United States, significant progress has been made in addressing media bias, with several noteworthy companies offering valuable methods and information. Platforms such as Ad Fontes Media[1], AllSides[2], The Factual[3], and Ground News[4] are well-known examples that study the issue of media bias, providing informative resources for news readers. Ad Fontes Media[2] utilizes expert annotators to meticulously assess news outlets, creating an interactive media chart focusing on two fundamental axes: reliability and political stance. AllSides, on the other hand, relies on expert annotators and community feedback to classify news outlets, offering a static media chart primarily centered around the political stance axis[48]. The work undertaken by these companies is particularly influential for this research, as it highlights the need for a similar platform addressing Portuguese news. In Chapter 3, a comprehensive study of these platforms is conducted, laying the groundwork for the proposed prototype in this research. By drawing upon successful and proven ideas from these influential websites, it is anticipated that

[1] adfontesmedia.com
[2] allsides.com
[3] thefactual.com
[4] ground.news

solid guidelines can be formulated for the development of a similar platform catering to Portuguese news.

2.5 Key Findings

Media bias has proven a harmful problem in today's society and has alerted the investigation by many researchers and research areas from social sciences to computer science. Typically, the social sciences analyze bias and try to detect instances of it using manual approaches. These approaches, however, require significant expertise and are time-consuming. In computer science, the NLP subject is mainly interested in the media bias problem as it has shown capable of solving similar tasks such as sentiment analysis. Thus, automated approaches to identify media bias use NLP techniques that try to detect a specific form of media bias – media bias by word choice and labeling.

Automated approaches are becoming more popular and a topic of investigation due to the recent advancements in the computer science subject brought by Transformers and BERT. Recent studies have shown that BERT-based models can accurately identify media bias occurrences and have opened the research gap to replace previous conventional techniques with state-of-the-art models. However, automated approaches suffer from the substantial drawback of the lack of gold-standard datasets. Many investigators have led projects to develop gold-standard datasets in the media bias area to tackle that issue.

2.6 Summary of the Chapter

As the news serve as the principal source of information for many people, news also impacts public opinion. News often shows slanted coverage of events resulting in the so-called media bias. In specific coverage situations, media bias can significantly impact the outcome of events or public decision-making, as shown by studies in presidential elections. However, the slanted coverage is not necessarily harmful and intentionally rendered as news results from an extensive creation process. Many definitions of media bias have been proposed over the years. However, in each investigation thesis, researchers select the most applicable definition for research as media bias is a broad research subject.

The creation process is defined in three phases, from the occurrence of an event to the actual consumer perception of the news reported. The first step is the gathering event, where the information is gathered, the event selected, and sources established. The next step is the actual writing of the news article. That is the most influential phase in the present research. In this phase, authors face word choice and labeling questions that might lead to a slanted selection toward a specific concept. The last stage is the presentation stage. In this stage, writers choose which pictures they will include and how they will caption them and allocate the size of the article (in the case of a physical newspaper).

The writing style of news articles is responsible for the word choice and labeling form of bias appearance. Word choice and labeling are concerned with how a concept is labeled. The research showed that people use semantic notions to categorize concepts and might perpetuate bias. Likewise, bias occurs by carefully selecting words near a concept to portray an unrealistic idea of a concept.

For many years the social sciences have analyzed media bias and developed approaches to identify them. The techniques are manual and require significant effort and expertise. The method usually relies on a codebook describing definitions, rules, examples, and directives coders should use to annotate information. This exercise is called content analysis and has a deductive and inductive approach. In the deductive method, the coders have an a priori codebook and annotate a corpus. In an inductive practice, the coders read the corpus without having the codebook directives. The objective is to find patterns or to devise the codebook only by knowing the research question and hypotheses.

In contrast, automated approaches have demonstrated remarkable efficiency in recent years, especially with the birth of BERT. Numerous researchers have been adapting and enhancing previous detection methods by leveraging this advanced technology, significantly improving various tasks. The ongoing nature of this work reflects the recognition that there are ample opportunities for further advancements in the field of media bias across multiple dimensions.

Regarding related work on media bias platforms, this study identifies two distinct types: fact-checkers and general bias identification platforms. Fact-checkers have gained popularity in Portugal, with numerous news outlets establishing dedicated fact-checking sections. However, there remains a noticeable gap in the Portuguese media landscape regarding platforms focused on identifying and disseminating information about media bias. Unlike the United States, where several companies actively work in this direction, no equivalent efforts exist in Portugal. This void presents a clear opportunity for this research to contribute by addressing and filling this gap in the Portuguese context.

Chapter 3

News Articles Framing Approach

3.1 Introduction 17
3.2 Identification of Axes of Interest 18
3.2.1 Reliability 19
3.2.2 Political Stance 19
3.2.3 Objectivity 20
3.2.4 Readability 21
3.3 Framing Approach 21
3.3.1 Datasets 22
3.3.2 Data Collection 23
3.3.3 Task 27
3.4 Prototype 30
3.5 Summary of the Chapter 34

This chapter elucidates the method employed in framing media bias and a general review of the underlying issue. The procedure for coming up with a solution is then made clear, along with the technique used to identify the various dimensions of media bias. Furthermore, it formulates a high-level strategy for addressing the axes. It ends with presenting a prototype specifically designed to resolve the intricacies of the problem at hand.

3.1 Introduction

In the contemporary landscape of our rapidly advancing digital era, the media plays a fundamental role in shaping public perceptions and attitudes. However, the escalating concern surrounding media bias has arisen due to its potential to disseminate inaccurate or incomplete information, significantly influencing individuals and society. Amidst the manifold manifestations of media bias, one particularly subtle yet pervasive approach strategically uses framing strategies through word choice and labeling. Framing entails deliberately presenting or structuring media content to mold the audience's perception of a particular issue or event. This section provides an in-depth overview of the proposed solution to detect media bias facilitated by word choice and labeling, thereby addressing the research questions outlined in Section 1.3.

The first research question pertains to employing natural language processing (NLP) techniques, as highlighted in the literature review. It has been demonstrated that leveraging these techniques can effectively uncover media bias within news articles.

The previous reinvigorates the second research question: "Which slants and attributes of media are instrumental in assisting news consumers in comprehending the nuances of media content?". In response to this query, this chapter delineates four pivotal axes believed valuable in aiding news consumers' understanding of media, accompanied by comprehensive explanations. Subsequent sections delve into a high-level approach to frame each axis and present effective methodologies for their detection. Finally, the chapter concludes by introducing a system prototype designed to visually depict these axes, thereby offering news consumers a tangible tool for enhancing their comprehension of media dynamics.

3.2 Identification of Axes of Interest

Media bias has become an intrinsic component of contemporary society, subtly influencing individuals' news consumption habits in ways often unnoticed. Individuals tend to lean towards news sources that align closely with their personal preferences, inadvertently reinforcing their biases. For instance, someone with a left-leaning political inclination is likelier to choose news outlets that portray left-wing parties favorably, even if it means neglecting coverage of other political factions. The human brain instinctively generates these preferences and expertly filters information that aligns with personal beliefs. This phenomenon is not limited to politics; individuals employ preconceived criteria to evaluate various aspects of their lives. For instance, people consider screen size, memory capacity, camera quality, and more factors when purchasing a new smartphone. Each of these criteria assumes a value that serves as a label to guide the decision-making process. Once the labeling is established, the potential buyer can discern which options align best with their preferences and requirements.

This section aims to break down the intricate criteria supporting the human brain's selection process regarding news articles. The literature review shows that significant research contributions have emerged, shedding light on detecting and identifying various forms of bias. These enclose not only the detection of fake news but also the exploration of racial discrimination, gender bias, political bias, and other related areas [42, 14, 4, 49]. Moreover, it is worth noting that companies specializing in the classification of American media bias [2, 48, 50, 51] have made notable strides in identifying additional criteria that aid in discerning biased articles. These criteria encompass evaluating source and website quality, analyzing writing tone and style, examining the presence and use of quotes, and other relevant factors. Such comprehensive approaches further enhance the understanding of the multifaceted nature of media bias and its impact on news consumption.

Building upon these insights, this section delineates four overarching axes – reliability, political stance, objectivity, and readability – that encapsulate and generalize the criteria mentioned earlier. These axes serve as a framework to empower news consumers with the necessary tools and knowledge to comprehend the articles they encounter. By providing a systematic and comprehensive understanding of media bias, individuals can navigate the news landscape more effectively and make informed judgments about the information they consume.

3.2.1 Reliability

Reliability constitutes a key aspect in evaluating the trustworthiness of a media outlet. A reliable media outlet is characterized by its commitment to the principles of journalism, consistently producing fair, accurate, and unbiased content. However, the digital era has given rise to challenges such as the rampant spread of false and misleading information. The perception of reliability regarding a news outlet holds immense significance in maintaining credibility and reputation [22, 6], as readers prioritize access to reliable information. Unfortunately, readers are often unaware of unreliable news outlets or articles, making them susceptible to the influence of misleading beliefs [9].

In the context of framing reliability, various natural language processing (NLP) techniques, such as sentiment analysis, topic detection, and named entity recognition (NER), play instrumental roles. Unreliable news sources frequently focus on a limited range of topics and exhibit consistent entity distribution within those topics. In conjunction with state-of-the-art text classification techniques, leveraging these NLP fields has demonstrated promising results [52, 53]. However, it is essential to note that in some instances, it may be necessary to incorporate additional information and methodologies beyond the scope of natural language processing to achieve even more accurate results. For example, The Factual [50] evaluates other news reports from the same author to classify reliability.

Empowering readers to assess the reliability of a news article before reading it enables consumers to make informed choices based on reliability scores, potentially mitigating their exposure to fake, unreliable, and misleading information. By providing readers with tools and indicators of reliability, it becomes possible for individuals to navigate the complex media landscape more effectively and make informed decisions about the news they consume.

3.2.2 Political Stance

Political stance encompasses the political orientation or alignment adopted by a media outlet or reflected in a news article. Media bias associated with political stance refers to a deliberate pattern favoring or promoting particular political orientations, ideologies, beliefs, or principles. It entails strategically choosing stories to relate to, framing them, and emphasizing certain aspects to support a particular political viewpoint.

Political bias within media outlets can take on various forms, each with distinct manifestations. One particularly prevalent form is observed in the framing of news articles, where specific political parties or ideologies are portrayed favorably while simultaneously unfavorably presenting opposing parties or ideologies. This deliberate framing subtly influences readers' perceptions and significantly impacts the narrative surrounding political events. By strategically offering information in a biased manner, media outlets have the potential to shape public opinion and steer the discourse on political matters.

Existing research has demonstrated the effectiveness of text classification techniques in identifying political bias [54, 55]. However, it is essential to note that most previous efforts have focused on social media users [55] and the broader scope of news media without explicitly delving into the domain of news articles. This highlights the need for more targeted investigations in this specific domain.

In the context of the Portuguese language, the issue becomes even more constrained due to the absence of publicly available news datasets that provide political stance or bias labeling, akin to platforms like AllSides[1] or Ad Fontes Media[2]. This scarcity of resources poses a significant challenge for researchers seeking to explore political bias in Portuguese news articles. Consequently, there is a pressing need to develop appropriate datasets or labeling sources to facilitate further studies in this area and enable a comprehensive understanding of political bias within Portuguese news outlets.

Finally, political bias has been demonstrated to be highly significant in shaping national and international events, such as elections. News consumers must adopt a critical mindset when accessing news as they are prone to various forms of bias, including political bias. However, such adoption requires an informed and discerning approach to news consuming and cultivating media literacy. By leveraging technology to classify political stances automatically, consumers can gain valuable insights that would otherwise necessitate significant effort and attention to achieve through traditional means of media analysis.

3.2.3 Objectivity

Objectivity, in the context of news articles, refers to the absence of opinionated concepts and terms. The subjectivity analysis aims to identify the presence of "private states", opinions, evaluations, beliefs, and speculations within the text [56, 57]. When selecting news articles, readers generally prefer factual reporting that presents information without bias or personal viewpoints. News writers may intentionally introduce their own opinions into their reports, sometimes subtly or inadvertently influencing the beliefs of news consumers. Accurate reporting, free from subjectivity,

[1] allsides.com - balanced news via media bias ratings for an unbiased news perspective.
[2] adfontesmedia.com - media watchdog organization.

allows users to understand events better as it presents information directly without the author's interpretation and subjective language.

In subjectivity analysis, text classification tasks have proven effective in discerning objective and subjective text, given the heavy reliance on language nuances. However, augmenting these tasks with additional features, such as assessing affective words, dynamic adjectives, semantic adjectives, and more [58], can further enhance performance. These other features provide a deeper understanding of the text's nuances and aid in distinguishing between objective and subjective elements.

By employing techniques that differentiate between objective and subjective content, news consumers can make more informed decisions about the articles they read. Objective reporting fosters a clearer understanding of events, unclouded by the personal interpretations and biases of the authors.

3.2.4 Readability

The readability of a news article is not necessarily a bias but an intrinsic attribute of the writing style. Considering this, it is difficult to relate this axis with the others as the others fall in the scope of existing biases in the news. Nevertheless, the axes aim to offer meaningful information to news consumers that can be used to understand the news they are consuming. News consumers may have personal choices according to the ease with which a text can be read.

Generally, readability ensures that news articles are presented clearly and concisely, easing the readers' job to acknowledge the information. Well-structured sentences, appropriate choice of vocabulary, and coherent paragraphs facilitate consumers understanding. These factors take part in the calculus of some readability indices, such as the Flesch–Kincaid readability test, Coleman–Liau index, or the automated readability index (ARI). Yet, some investigations have shown that these rule-based readability indices cannot fully capture writing quality [59] being occasionally outperformed by other techniques such as BERT models for classification.

Text from diverse media sources is frequently challenging to comprehend, resulting in less-than-optimal message comprehension. This is where readability detection can help, as it allows readers to assess the difficulty level of a piece of text and decide whether or not to invest their time and effort in reading it. Overall, readability promotes transparency, inclusivity, and comprehension, enabling readers to understand the media better and make well-informed decisions in an information-rich landscape.

3.3 Framing Approach

This section presents the overarching approach for framing news within the identified axes. The method follows a systematic process that involves searching for relevant datasets, collecting data when existing datasets are insufficient, exploring the collected data, developing a machine learning model (preferably leveraging existing pre-trained models), and evaluating the model's performance.

3.3.1 Datasets

The dataset search process targeted topics aligned with the pre-defined axes. The search was narrowed to English, European Portuguese, or Brazilian Portuguese datasets to ensure relevance and suitability. Moreover, there was a preference for expert-annotated datasets, meaning they had been curated and reviewed by domain experts to ensure high quality and accuracy. This approach aimed to gather datasets that were not only linguistically appropriate but also enriched with expert knowledge, maximizing the value and reliability of the collected data.

Table 3.1 presents the most notable datasets that have been identified. These datasets have been instrumental in conducting valuable research on topics such as media bias [38] and objectivity detection [60]. However, upon conducting a thorough dataset search, it became apparent that datasets specifically tailored to the identified axes of interest are limited.

Table 3.1: Gathered datasets for the classification of the media axes.

Name	**Related Axis**	**Samples**	**Language**
BABE [38]	Objectivity	3 700	English
Portuguese Parliamentary Minutes	Political stance	5 713	European Portuguese
Portuguese Fake News [61]	Reliability	103 966	European Portuguese

Despite the significance of these datasets in related research, their applicability to the identified axes remains scarce. This scarcity highlights the need for further efforts to collect or develop datasets that align more closely with the specific axes under investigation. Such efforts would significantly advance research in these areas and enable more comprehensive analyses and evaluations.

The BABE dataset, introduced in the study by Spinde et al. [38], is a collection of news sentences annotated by experts. This dataset builds upon a smaller expert-annotated dataset known as MBIC [42], which consists of 1 700 annotated sentences. In the aforementioned study, the researchers utilized the MBIC dataset's 1 700 sentences and an additional set of 2 000 sentences, resulting in two distinct subgroups of annotations. The annotation process involved a group of eight annotators who first annotated the original MBIC sentences. Subsequently, out of these eight annotators, five were selected to annotate the additional 2 000 sentences. This collaborative effort resulted in a total of 3 700 meticulously annotated news sentences available in the BABE dataset. Within the BABE dataset, the individual sentences from news articles have been annotated across three distinct categories: political orientation (left, center, right), bias label (biased, non-biased), and opinion label (factual, expresses writer's opinion, somewhat factual but also opinionated).

The political orientation category is automatically assigned by utilizing the AllSides political orientation as a reference. On the other hand, the remaining two categories are manually annotated, ensuring a detailed and accurate labeling process.

To incorporate the BABE dataset into the research for this dissertation, a translation process was necessary as the dataset was not initially available in the target language. For this purpose, the DeepL Translate API[3] was utilized. Unlike Google Translate, which does not differentiate between European Portuguese and Brazilian Portuguese, DeepL offers the advantage of allowing the selection of the target language specifically for European Portuguese.

The Portuguese Parliamentary Minutes dataset, available at [4], offers a collection of parliament interventions extracted from the Plenary sessions held in the Portuguese Parliament. The dataset was created by parsing the PDF minutes of these sessions, enabling the extraction of interventions made by the parliament members. It is important to note that the collected data within this dataset required cleaning and curation. In its raw form, the dataset contained inconsistencies or noise that could impact its usability for certain analyses. A curation stood to address these concerns, which involves eliminating excessively long and very short interventions consisting of only a few words. This curation process results in a final dataset containing 5 713 relevant and suitable samples for further analysis. It is important to note that the Portuguese Parliamentary Minutes dataset is not a dataset within the domain of news articles. The dataset represents texts that were said by deputies in the Portuguese parliament. However, due to the lack of datasets for this axis, this is the best suitable option to pursue political stance identification.

In 2021, Moura et al. [61] collected articles from Portuguese fake news websites. They used a similar approach to one of the data collection tool, explained in the next section. The data was collected between 2017 and 2020 from *Bombeiros24*, *JornalDiario*, *MagazineLusa*, *NoticiasViriato*, and *SemanarioExtra*.

3.3.2 Data Collection

As previously mentioned, there is a significant scarcity of data available in Portuguese that is specifically related to the identified axes discussed in Section 3.2. In some cases, relying on Brazilian Portuguese data or translating existing datasets to Portuguese may not be a viable solution. To illustrate this, let's consider a dataset labeled with political orientation. The political landscape in Brazil or any English-speaking country may differ significantly from the political panorama in Portugal. Therefore, it would be incorrect to translate or utilize a Brazilian Portuguese dataset, as it would neglect the previously mentioned assumption. In such cases where gathering new data was required, this work used a news collection tool to collect and curate such data. This section describes the data collection tool's capacities and limitations and shows how it can be highly helpful in grouping data.

[3] deepl.com/docs-api

[4] github.com/afonso-sousa/pt_parliamentary_minutes

The data collection tool employed in this study consists of a series of automated scripts written in Python. These scripts are designed to extract data from news websites, focusing on news articles. The tool utilizes the Arquivo.pt[5] [62] Web archive as its primary resource for crawling news web page URLs based on a given query. Subsequently, the scripts use Python libraries to download and parse the content of the web pages, enabling efficient data retrieval.

The flowchart depicted in Figure 3.1 visually represents the data collection process. The process begins by taking an input configuration file as its starting point. Within this file, a list of websites is specified. Afterward, the process proceeds to fetch the URLs of the articles from the designated websites mentioned in the configuration file. Once the article URL is retrieved, each web page's content is downloaded. In the final stages of the process, each article is parsed, extracting relevant information from its content. The gathered data, including various attributes of the articles, is then saved in a JSON file. This file is a structured storage format, enabling easy access and manipulation of the collected data for further analysis or processing.

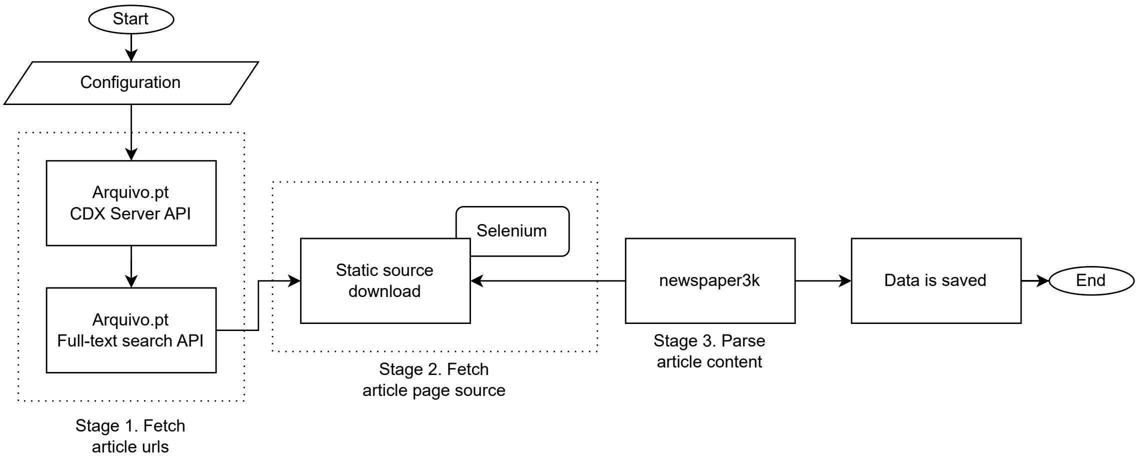

Start
Configuration
Arquivo.pt CDX Server API
Arquivo.pt Full-text search API
Stage 1. Fetch article urls
Selenium
Static source download
Stage 2. Fetch article page source
newspaper3k
Stage 3. Parse article content
Data is saved
End

Figure 3.1: Flowchart of the data collection process.

The following link is the GitHub repository of the data collection tool: github.com/luist18/article-scraping.

3.3.2.1 Configuration

As previously mentioned, the configuration of the data collection process is accomplished through a dedicated configuration file. This file, formatted as JSON, must adhere to a predefined syntax to ensure proper functioning. Various parameters can be customized to suit specific requirements within this configuration file.

The configurable parameters encompass the following key aspects:

- **Websites**: The configuration file allows for the specification of the websites to be included in the data collection process. These websites are typically chosen based on their relevance to the research or the desired scope of the data collection.

[5] arquivo.pt - non-profit Portuguese Web archive service.

- **Queries**: The configuration file includes specific queries that will be used to retrieve relevant articles from the designated websites. These queries serve as search terms or criteria to narrow down the focus of the data collection.
- **Date Range**: The configuration file also provides the flexibility to define a specific date range for the search. This allows for the retrieval of articles published within a particular timeframe, which can help capture data from specific periods of interest.
- **Output File**: Lastly, the configuration file allows for the specification of the desired name for the output file. After processing and organizing, the collected data will be saved in a JSON file with the designated name, facilitating easy identification and future reference.

Appendix A.1 shows an example configuration file to search the terms "Brexit" and "European Union" in The Guardian[6] website from January 1st, 2019 to December 31st, 2020.

3.3.2.2 Stage 1: Articles URL fetching

The initial stage of the crawling process involves retrieving the URLs of articles that meet the specified criteria in the configuration. This process consists of two components: news source website URL fetching using the Arquivo.pt CDX Server API[7] and article URL fetching using the Arquivo.pt full-text search API[8].

The CDX Server API enables the retrieval of all URL versions of a specific URL. For instance, the current URL for the Portuguese newspaper Observador's website is observador.pt. However, in the past, it used to be www.observador.pt. Although the website now automatically redirects the browser to the new URL, which is essentially the same, capturing these changes is crucial. This is because, in the second component of this process, the full-text search API is highly sensitive to even minor variations like these. It indexes different content under the URLs observador.pt and www.observador.pt. Appendix B.1 demonstrates an example of the CDX Server API response. The outcome of this segment is an extended list of website URLs from the websites defined in the configuration file.

The next component, article URL retrieval, relies on the various versions of the URLs obtained from the previous step. This retrieval process utilizes the full-text search API, an interface for searching and accessing preserved web content and associated metadata indexed by Arquivo.pt. The API offers numerous customizable parameters, including the query term, date range, and a list of specific websites to search within. In this particular component, each query defined in the configuration for each website URL from the previous step is explored within a specified date range. An example of the API response to such a query can be seen in Appendix B.2, which demonstrates the results obtained when searching for "Brexit" in articles from The Guardian, covering the period from January 1st, 2019, to December 31st, 2020.

[6] theguardian.com/uk - The Guardian's UK Edition website.

[7] github.com/arquivo/pwa-technologies/wiki/URL-search:-CDX-server-API

[8] github.com/arquivo/pwa-technologies/wiki/Arquivo.pt-API

3.3.2.3 Stage 2: Article Page Source Fetching

Stage 2 focuses on the crucial task of downloading the web page source from the provided URLs. The source comprises the raw HTML content rendered on the articles' pages. While this process may initially appear simple, it requires careful consideration due to the various methods available for acquiring web page content.

One such technique involves using the command-line tool `curl`, which permits the retrieval of web page content by running a command like `curl www.google.com` to fetch the content of www.google.com. The Python Software Foundation, the organization behind Python, has also developed an open-source HTTP Python library named `requests`[9]. This library offers functionalities to fetch web page content and manipulate HTTP requests, driving it into a valuable instrument.

The methods mentioned earlier are effective for retrieving content from server-side rendered websites. These are websites where the content is generated on the server and then delivered to the user. However, these methods encounter limitations when dealing with client-side rendered websites. Unlike server-side rendered websites, client-side rendered websites do not fully generate the content initially. Instead, they rely on scripts to fetch data and dynamically render it in the user's browser as they navigate the website.

This distinction poses a challenge for the previously described methods, as they are unaware of these scripts and can only retrieve the content directly provided by the server during the initial request. Consequently, they may not capture the complete content or dynamically rendered elements from client-side rendered websites.

To address this challenge, when the content cannot be obtained using the `requests` library, this stage employs another Python library known as `selenium`[10]. This library allows the simulation of a web browser, enabling the script to wait for the page to be dynamically rendered before retrieving the page source. By utilizing `selenium`, the script can interact with the client-side generated website, execute the necessary scripts, and capture the fully rendered content. However, it is worth noting that this approach typically exhibits slower performance than the method used for server-side generated websites. The additional steps in simulating a web browser and waiting for dynamic rendering increase processing time.

As a result of this stage, a collection of page sources corresponding to the URLs gathered in the first stage takes shape. It is worth noting, however, that when both the `requests` and `selenium` libraries fail to get the page source, the article associated with the failed URL is ignored and removed from the final set of page sources. This ensures that only page sources that have been correctly obtained are evaluated for further processing.

[9] github.com/psf/requests

[10] selenium-python.readthedocs.io

3.3.2.4 Stage 3: Articles Parsing

The concluding stage of the data collection process involves parsing the gathered page sources. This parsing task uses a specialized article extraction and curation library called `newspaper3k`[11]. By providing the HTML content of a webpage as input, `newspaper3k` can extract various attributes associated with an article.

One of the primary functions of `newspaper3k` is to cleanse the text extracted from a webpage, particularly when it is accompanied by noise, such as advertisements. The library performs this cleaning process to ensure the extracted content is as accurate and relevant as possible. Moreover, `newspaper3k` can retrieve additional information about the article, including the authors, title, main content, publication date, and tags, among other attributes. This comprehensive feature set makes `newspaper3k` an indispensable tool for extracting and curating article data from web pages.

Upon completion of this stage, the result is a collection of parsed articles, each possessing a set of essential attributes, including:

- **Title**: The article's title concisely summarizes its subject.
- **Content**: The article's main body, comprising the informative text and details presented.
- **Date**: The article's publication date, indicating when it was initially released or made available.
- **Tags**: Relevant keywords or tags associated with the article, facilitating categorization and topic identification.
- **URL**: The web address (URL) pointing to the article's source allows easy reference and access.

These attributes provide valuable information about each parsed article, enabling further analysis, organization, and utilization of the collected data.

3.3.3 Task

Various methodologies have been utilized in textual tasks, exhibiting a wide range of diversity. A significant breakthrough occurred in 2017 with the emergence of transformer models, as demonstrated by Vaswani et al. in their work on attention mechanisms [36]. This milestone was followed by the deployment of the BERT model by Devlin et al. in 2019 [37], further cemented the effectiveness of transformer-based models. Consequently, these models have consistently achieved remarkable state-of-the-art performances across numerous tasks.

The objective of addressing each axis is accomplished by fine-tuning pre-trained transformer models for news article framing. This section provides an overview of a framework designed to analyze each axis and construct a model specifically tailored for framing news articles.

[11] github.com/codelucas/newspaper

The procedure involves a sequential approach that includes searching for suitable datasets, creating one if no viable options are available, thoroughly exploring the data, developing a model, and conducting a comprehensive evaluation of its performance. The process further developed is suitable for addressing the topics of news reliability, objectivity, and political stance. However, when it comes to the readability axis, it will be tackled with the existing indexes that do not require training.

The initial step in addressing each axis involves the collection of appropriate datasets. While searching for datasets, it was discovered that no reliable datasets in European Portuguese were available, as indicated by the findings presented in Table 3.1. To overcome this limitation, Section 4.2.3 introduces a new dataset of Portuguese news articles classified into reliable and unreliable categories. It is important to note that each axis is not limited to a single dataset. This work demonstrates the utilization of multiple datasets for training a model. One viable strategy is employing a dataset to benchmark the model's performance. Alternatively, merging two or more datasets can be helpful when data scarcity is a concern.

Once a collection of datasets has been gathered, exploring the data to understand its structure and composition is fruitful comprehensively. This exploration process involves carefully analyzing the data and identifying any samples deemed unrelated, noisy, or duplicated, thereby allowing for their removal. Furthermore, data exploration often leads to valuable insights, unveiling correlations between attributes and the target label and unearthing beneficial facts about the attributes themselves. These insights can prove highly valuable in informing subsequent analysis and model development stages.

Upon completing the exploration and preparation of the data for model training, a meaningful consideration arises with data augmentation. Data augmentation refers to the process of enriching textual data using various techniques. Some standard methods include synonym replacement, text paraphrasing, contextual word embeddings (replacing words with contextually similar words using pre-trained word embeddings), back-translation, etc. Section 4.3.2 explores a specific data augmentation technique in the Portuguese Parliamentary minutes' dataset context. This section provides a detailed examination of how data augmentation can be applied to enhance the dataset for improved model performance.

The next step is building a model for the dataset. This dissertation uses multiple pre-trained transformer models that support the Portuguese language. As mentioned, the aim is to fine-tune existing models and suit each axis's needs. There are several Brazilian Portuguese and European Portuguese models and multilingual models. BERTimbau [63] and Albertina-ptbr [64] are examples of models trained in Brazilian Portuguese, Albertina-ptpt [64] is an example of a European Portuguese model, and XLM-RoBERTa [65] is a multilingual model. Table 3.2 show these models' available versions, their base architecture, and the number of parameters.

To build a classification or regression model from BERT, it is necessary to take advantage of the pooler output of the BERT architecture. The purpose of the pooler layer is to generate a fixed-size representation, often referred to as the "pooled output", that summarizes the input sequence's

Table 3.2: Examples of Brazilian Portuguese, European Portuguese, and multilingual transformer-based pre-trained models.

Model	**Builds On**	**Parameters**	**Language**
BERTimbau$_{\text{BASE}}$ [63]	BERT	110M	Brazilian Portuguese
BERTimbau$_{\text{LARGE}}$ [63]	BERT	335M	Brazilian Portuguese
Albertina PT-PT$_{\text{BASE}}$ [64]	DeBERTa	100M	European Portuguese
Albertina PT-BR$_{\text{BASE}}$ [64]	DeBERTa	100M	Brazilian Portuguese
Albertina PT-PT$_{\text{LARGE}}$ [64]	DeBERTaV2	900M	European Portuguese
Albertina PT-BR$_{\text{LARGE}}$ [64]	DeBERTaV2	900M	Brazilian Portuguese
XLM-RoBERTa$_{\text{BASE}}$ [65]	RoBERTa	279M	Multilingual
XLM-RoBERTa$_{\text{LARGE}}$ [65]	RoBERTa	561M	Multilingual

contextual information. The pooler layer processes the entire sequence of hidden states from the last layer and produces a condensed representation of the input sequence. This pooled output typically represents the entire sequence in downstream tasks such as sentence-level classification or similarity tasks.

The process of constructing a model depends on the dataset and the nature of the target labels. For instance, consider a political orientation dataset where the task is to assign a numerical value between 0 and 5 to each sample, with 0 representing left-oriented and 5 denoting right-oriented. In this case, the problem can be classified as a regression problem, aiming to predict a continuous value.

However, in practice, datasets often have discrete labels rather than continuous ones. Consider the same dataset, but now the labels are annotated with three categories: left, center, and right. These categories can be converted into numerical values while maintaining the value range. However, assigning exact numbers to each sample can create ambiguity. For example, the label "left" may not always correspond to 0 but could be represented as 0.5 or 1.2 in some instances. To handle such scenarios where the target assumes discrete values, the problem is referred to as a classification problem. In classification, the objective is to assign a unique value or category to each sample based on its characteristics and features.

In summary, building a model depends on the dataset and the type of target labels. While regression is used when predicting continuous values, classification is employed for discrete label values. The labels initially represented as numbers can be transformed into discrete categories in the political orientation dataset, introducing classification as the appropriate problem-solving approach.

The choice of model evaluation metrics and loss functions varies depending on the problem being addressed. In regression problems, evaluation metrics commonly include the mean squared error (MSE) and mean absolute error (MAE). These metrics assess the degree of error between predicted and actual continuous values, providing insights into the model's accuracy in regression tasks.

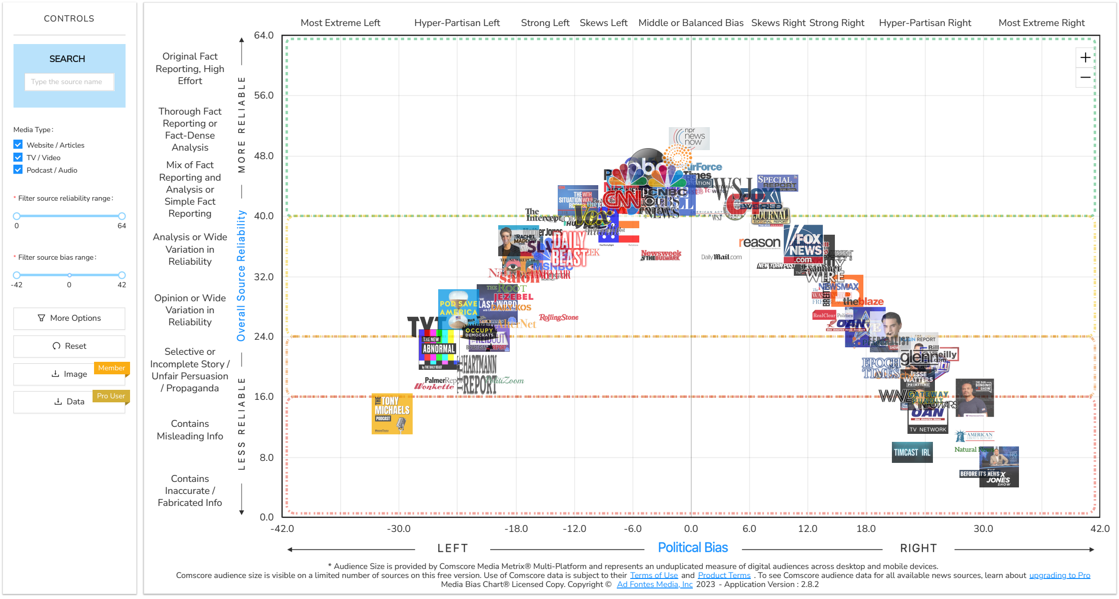
On the other hand, classification problems involve assigning samples to discrete categories. In such cases, accuracy is a widely used evaluation metric that measures the proportion of correctly classified samples. Additionally, the F-score, which combines precision and recall, is often employed to assess the model's performance across multiple classes. In terms of loss functions, for classification problems, the Cross-Entropy Loss is frequently utilized. This loss function evaluates the dissimilarity between predicted class probabilities and the actual class labels. The model learns to generate accurate probability distributions over the different classes by minimizing the cross-entropy loss.

It's important to note that these are just some examples of commonly employed metrics and loss functions, and the specific choice can vary depending on the nature of the problem and the desired outcome. The selection of appropriate evaluation metrics and loss functions plays a crucial role in assessing and optimizing the performance of machine learning models. Chapter 4 shows each model's loss function and evaluation metrics and a detailed explanation and reason for using them.

3.4 Prototype

This section emphasizes the importance of presenting the model's predicted data to news consumers in an interactive and simplified manner. This user-friendly approach enables news consumers to grasp and interpret the information generated by the models quickly. By providing an interactive interface or visualizations, the insights derived from the model can be effectively conveyed, enhancing the user experience and facilitating informed decision-making. To develop a prototype specification, it was imperative to leverage existing work and gather ideas that have proven to be effective. In this regard, thorough consideration was given to reputable sources such as Ad Fontes Media, All Sides, and The Factual. These organizations are known for their expertise in news media analysis and have provided valuable insights and methodologies that can significantly contribute to developing a robust prototype.

Ad Fontes Media has a manual process of annotating United States news and news sources. The annotations are done according to two categories: reliability and political bias. Ad Fontes Media presents an interactive media bias chart with annotated data. The chart is a 2D chart where each news article, TV video, and podcast is displayed. The reliability values range from 0 to 64. Scores surpassing 40 are typically considered favorable, while scores below 24 tend to raise concerns. Scores falling within the 24-40 range encompass a diverse spectrum of possibilities. Some sources within this range may exhibit a prevalence of subjective viewpoints and analytical content, while others may display significant disparities in reliability across their articles. The political bias scores assigned to articles and shows are measured on a scale ranging from -42 to +42. In this scale, higher negative scores indicate a more substantial left-leaning bias, higher positive scores indicate a stronger right-leaning preference, and scores closer to zero suggest minimal bias, a balanced perspective, or a centrist bias. Figure 3.2 shows the interactive media bias chart on the Ad Fontes Media website. Each media outlet is represented in the chart with the mean value of each axis.



CONTROLS

SEARCH

Type the source name

Media Type:
☑ Website / Articles
☑ TV / Video
☑ Podcast / Audio

* Filter source reliability range: 0 – 64

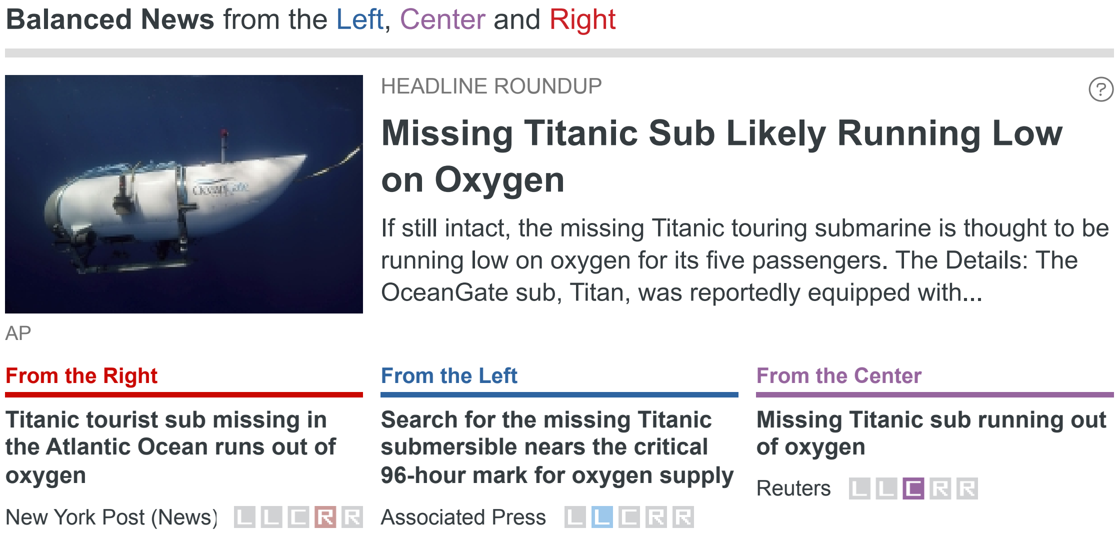
* Filter source bias range: -42 – 0 – 42

More Options
Reset
Image
Member
Data
Pro User
Most Extreme Left
Hyper-Partisan Left
Strong Left
Skews Left
Middle or Balanced Bias
Skews Right
Strong Right
Hyper-Partisan Right
Most Extreme Right
Original Fact Reporting, High Effort
Thorough Fact Reporting or Fact-Dense Analysis
Mix of Fact Reporting and Analysis or Simple Fact Reporting
Analysis or Wide Variation in Reliability
Opinion or Wide Variation in Reliability
Selective or Incomplete Story / Unfair Persuasion / Propaganda
Contains Misleading Info
Contains Inaccurate / Fabricated Info
MORE RELIABLE
Overall Source Reliability
LESS RELIABLE
64.0
56.0
48.0
40.0
32.0
24.0
16.0
8.0
0.0
-42.0
-30.0
-18.0
-12.0
-6.0
0.0
6.0
12.0
18.0
30.0
42.0
LEFT
Political Bias
RIGHT
+
−

* Audience Size is provided by Comscore Media Metrix® Multi-Platform and represents an unduplicated measure of digital audiences across desktop and mobile devices.
Comscore audience size is visible on a limited number of sources on this free version. Use of Comscore data is subject to their Terms of Use and Product Terms . To see Comscore audience data for all available news sources, learn about upgrading to Pro
Media Bias Chart® Licensed Copy. Copyright © Ad Fontes Media, Inc 2023 - Application Version : 2.8.2

Figure 3.2: Ad Fontes Media interactive chart [2].

AllSides annotates media news and outlets according to their political bias. It uses various identification methods. They entrust editorial reviews, blind bias surveys, community feedback, and third-party data. The editorial reviews are conducted by a panel of six to nine reviewers with diverse political affiliations. These reviewers assess news reports for signs of bias, discuss their findings, and ultimately assign a bias rating. The panel examines explicitly aspects such as slant, spin, sensationalism, and story selection to identify common forms of media bias. Using tools like the Wayback Machine, they analyze the outlet's homepage, headlines, recent articles, images, and other content spanning up to six months. Each reviewer independently assigns a numerical rating. These numerical ratings are then combined to calculate a weighted average. The blind bias surveys consist of surveys directed to average Americans from all political spectrum that individually classify news. The survey is considered blind as all branding and identifying details are removed so respondents are not affected by preconceived ideas of an outlet's bias. On the AllSides website, the community can "agree" or "disagree" with the bias assigned to each media outlet. This way, the website can warn the users how many users agree or disagree with the given label. This method does not determine the bias rating of AllSides. However, it is a valuable indicator that the current bias rating may be inaccurate, prompting a more thorough examination. Finally, AllSides sometimes considers academic works, research, or analysis that appear transparent on the bias detection method. Figure 3.3 shows how AllSides shows related news from various media outlets. Typically, they present a headline summary with a report from the left, center, and right.



Balanced News from the Left, Center and Right

HEADLINE ROUNDUP

Missing Titanic Sub Likely Running Low on Oxygen

If still intact, the missing Titanic touring submarine is thought to be running low on oxygen for its five passengers. The Details: The OceanGate sub, Titan, was reportedly equipped with...

AP

From the Right

Titanic tourist sub missing in the Atlantic Ocean runs out of oxygen

New York Post (News) L L C R R

From the Left

Search for the missing Titanic submersible nears the critical 96-hour mark for oxygen supply

Associated Press L L C R R

From the Center

Missing Titanic sub running out of oxygen

Reuters L L C R R

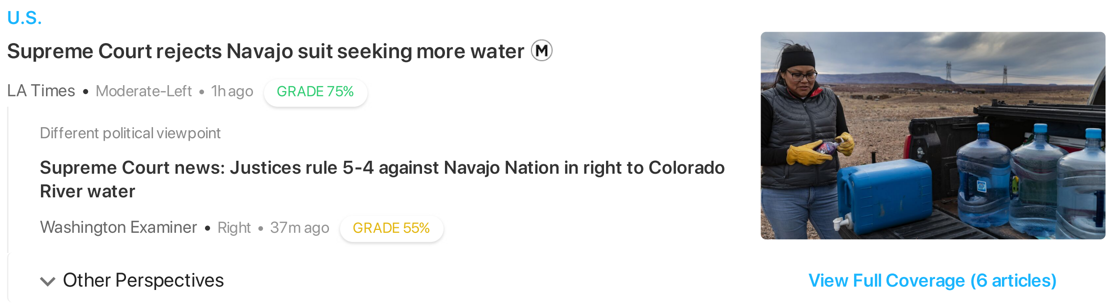
Figure 3.3: AllSides headline roundup with news about the missing Titanic sub.

The Factual utilizes an automated approach to assign a grade to each news article. This grade is accompanied by additional political context obtained from sources such as AllSides and Media Bias Fact Check[12], which provide political bias ratings. The Factual's grading system considers several factors: site quality, author expertise, source quality, and tone analysis. Site quality is evaluated based on the historical performance of the site in producing high-scoring articles. This metric considers the average quality of the articles published on the site over time. Author expertise is assessed by analyzing the author's previous writings. The evaluation finds the informativeness of their past reports and their level of knowledge in specific topics. For instance, an author who has written numerous articles might have a lower level of expertise in a particular topic than a more experienced writer who focuses on that specific area. Source quality is determined by considering factors such as source diversity, the presence of meaningful quotes, and the inclusion of unique links. These criteria help assess the reliability and credibility of the sources cited in the article.

At last, The Factual analyzes the tone of the article's text to detect whether it presents a neutral perspective or exhibits opinionated emotions. Considering these dimensions, The Factual aims to provide a comprehensive and nuanced assessment of news articles, incorporating site quality, author expertise, source quality, and tone analysis to deliver a more informed and contextualized grade. These metrics are then combined to produce a single percentage grade, representing the article's probability of being informative. Articles with grades exceeding 75% are considered highly likely to be informative, while grades below 50% indicate a lower likelihood of being informative.

[12] mediabiasfactcheck.com

Figure 3.4 portrays The Factual's coverage with multiple media outlets for the Supreme Court decision on the Navajo Nation suit in the right to Colorado River water.



U.S.

Supreme Court rejects Navajo suit seeking more water Ⓜ

LA Times • Moderate-Left • 1h ago GRADE 75%

Different political viewpoint

Supreme Court news: Justices rule 5-4 against Navajo Nation in right to Colorado River water

Washington Examiner • Right • 37m ago GRADE 55%

Other Perspectives

View Full Coverage (6 articles)

Figure 3.4: The Factual news coverage of the Supreme Court rejection against Navajo Nation.

The examples above offer valuable insights into the presentation of information for news consumers. Ad Fontes Media stands out with its distinct interface, featuring a two-dimensional chart that provides a novel and comprehensive interpretation of the media landscape. However, it lacks certain features that allow users to access individual news articles directly without the need to click on their position in the chart, which may be inconvenient.

On the other hand, AllSides presents information on individual pages for each news report. While it only utilizes one dimension, the square-shaped marker immediately shows the users the political orientation of the report. This approach offers a different scenario where users can quickly understand the political leanings of news articles at a glance.

In contrast, The Factual adopts a four-dimensional approach. However, to simplify the experience for news consumers, it combines all these dimensions into a single representation. This streamlining allows for a more straightforward categorization process. Nonetheless, additional information becomes available when expanding the article's grade, providing users with more detailed insights.

Each approach provides different perspectives and user experiences in presenting information to news consumers, offering a range of options to suit individual preferences and information needs. The main takeaways from the examples described are representing information through interactive ways, such as charts (Ad Fontes Media); straightforwardly presenting information, allowing consumers to immediately understand what type of news they are looking at (AllSides); providing detailed information about all dimensions (The Factual).

Considering the insights gained from the examples above, the proposed prototype can be developed as a web platform to provide news consumers with information about news articles and media outlets. The website should incorporate the concepts of dimensions or axes, allowing users to observe the results for each dimension in individual articles. Additionally, users should be able to obtain an overview of all dimensions collectively. To facilitate interaction and understanding, the prototype should feature a two-dimensional chart for regression models, where each axis represents a specific dimension. This chart would let users quickly interpret the results and assess the article's characteristics.

On the other hand, for classification models, a radar chart would be more appropriate. Classification results often cannot be effectively represented on a linear scale, and a radar chart allows for a visual depiction of the various dimensions involved in the classification process, enhancing users' interpretation of the results. By incorporating these design elements, the prototype can provide an intuitive and user-friendly platform that empowers news consumers to access and comprehend the information interactively generated by the models.

3.5 Summary of the Chapter

This chapter tackles the issue by identifying and describing four important axes in news story framing. These are the dimensions of reliability, political stance, objectivity, and readability. Reliability is essential in determining a media outlet's reliability and credibility. It determines how much readers may trust the information provided by the publication. The axis of political stance questions news articles' position about the political spectrum. It identifies and characterizes news articles' political orientation or bias, helping readers understand their leanings and perspectives. Objectivity, however, identifies the lack of opinion in news articles. It stresses the importance of giving information factually, devoid of subject views, biases, or personal beliefs. Lastly, the readability axis focuses on how articles are written and evaluate their ease of comprehension for readers. By defining and addressing these four axes, this chapter lays the foundation for a comprehensive framework that enables the analysis and framing of news articles in a multidimensional manner, encompassing reliability, political stance, objectivity, and readability.

Subsequently, this chapter outlines two distinct methodologies designed to frame the identified axes. The first methodology leverages machine learning techniques, specifically utilizing transformer-based models. This approach involves training and fine-tuning these models to analyze and assess the different axes, such as reliability, political stance, and objectivity. However, in the case of readability, a different approach is taken, relying on predefined rules and readability index formulas to evaluate the readability of news articles. The machine learning methodology depends on the availability of suitable datasets for training the models. If no appropriate dataset is readily accessible, this chapter proposes an alternative solution. It demonstrates how data can be obtained by crawling the Portuguese internet archive, Arquivo.pt. By extracting data from this archive, valuable information can be collected, enabling the construction of a dataset for further analysis. After having a dataset, it shows the different strategies to approach a dataset. It gives an overview of the methods that are going to be used in Chapter 4.

Henceforth, the notion for the prototype of this work is defined as a web platform that incorporates the best working ideas from multiple consolidated platforms. It has been decided to display the information in two distinct ways: chart-wise and detail-wise. With charts, the users can interact with the data and gain knowledge about a general panorama more quicker. However, detailed information can also be helpful to identify specific cases or just the lean of a news article.

Chapter 4

Axes-oriented Automatic Classification

4.1 Introduction 37

4.2 Reliability 37

4.2.1 Related Work 37

4.2.2 Datasets 38

4.2.3 Exploration 39

4.2.4 Method 42

4.2.5 Evaluation 45

4.3 Political Stance 46

4.3.1 Related Work 47

4.3.2 Datasets 48

4.3.3 Exploration 50

4.3.4 Method 51

4.3.5 Evaluation 53

4.4 Objectivity 55

4.4.1 Related Work 55

4.4.2 Datasets 55

4.4.3 Method 56

4.4.4 Evaluation 56

4.5 Readability 58

4.5.1 Readability Indices and Formulas 58

4.6 Future Work 62

4.7 Summary of the Chapter 63

This chapter provides a broad understanding of the process undertaken by each axis to classify news articles. Each axis carefully elucidates its methodology, showcasing the steps, techniques, and resources to achieve accurate classification. Following the detailed exposition of the methods employed in each axis, an examination of future work is presented, promoting a discussion on the forward trajectory of this research topic.

4.1 Introduction

In the following sections, the models that have been developed for each axis are presented. These machine learning models were executed on a server with an Intel Xeon W-2295 @3.0GHz processor, two Quadro RTX 8000 GPUs, each with 48GB of RAM, and a total system RAM of 128GB. However, the models were run without GPU parallelism, only taking advantage of one GPU. This was because the high-processing computing unit was shared. All random seeds were set to a fixed value of 42 to ensure consistency and facilitate result comparison. Additionally, during training, early stopping was used with the patience of 3 epochs, whereby the training process would halt if there were no improvements in the validation loss. When the patience limit was reached, the model with the lowest validation loss was restored and considered for further evaluation. The validation occurred thrice per epoch at 33%, 66%, and 100% of the epoch steps.

4.2 Reliability

This work defines reliability as trustworthiness and responsibility in producing fair, accurate, and unbiased content. This section provides a thorough methodology for framing news articles on their reliability. An analysis of related work on Portuguese news is conducted, shedding light on existing research and approaches in this domain. This serves as a foundation for the subsequent steps. Next, the creation of a novel dataset comprising reliable and unreliable Portuguese news articles is explained. This process involves careful selection and curation to ensure the dataset represents diverse sources and perspectives. Once the dataset is established, it is explored to uncover meaningful insights. This exploration aims to identify patterns, trends, and distinguishing characteristics that can inform the classification of news articles based on their reliability. Subsequently, the methodology employed to classify news articles according to their reliability is presented. This methodology utilizes BERT-based models to assign a reliability label to each article in the dataset. Finally, the performance of the classification model is evaluated. This evaluation considers relevant metrics to measure the model's ability to classify news articles on their reliability accurately.

4.2.1 Related Work

Social science studies have explored reliability-related areas, such as online disinformation in the context of Portuguese elections [66]. Another study investigated malicious political activities by analyzing tweets during the 2019 Portuguese elections [67]. Furthermore, considerable research focused on detecting fake news in Portuguese with an automated method [61]. While these studies provide valuable insights, none directly address our work's specific objective: developing machine-learning techniques for detecting reliability in Portuguese news articles.

In their study, Baptista et al. [66] examined online disinformation during the 2019 Portuguese Elections. Their investigation delved into the social media impact and dissemination of fake news within Portugal. Their research findings unequivocally confirmed fake news and misinformation within the country. To accomplish this, the researchers scrutinized the Facebook pages of prominent Portuguese newspapers and identified fake news and unreliable sources on Facebook pages through the findings of Pena [68].

Similarly, Ramalho [67] also explored the 2019 Portuguese Elections with a different focus. His study centered on analyzing the Portuguese Twittersphere, aiming to detect malicious political activity. In addition, Ramalho's work identified Portuguese websites responsible for disseminating fake news and unreliable content.

In 2021, Moura [61, 69] studied Portuguese fake news using machine learning and computational forensic linguistics approaches. His work is perhaps similar to the work developed in this axis. He collected data from multiple fake news websites to build a machine-learning model that topped with an accuracy of 0.99 and a macro averaged F_1 score of 0.97. Apart from his work, no prior research on machine learning techniques on Portuguese news reliability has been done that has available results to perform a comparison. Thus, his work will be used to benchmark the model performances in the following sections, as it is the only available result to compare.

4.2.2 Datasets

The training dataset combines data collected by Moura [61] and the data specifically gathered for this research. The data collection tool was utilized to gather unreliable and reliable news articles from specific websites to augment the training dataset and make it more comprehensive. A significant addition to the existing dataset was the inclusion of satire news articles obtained from *Inimigo Público*. Sixteen unreliable websites were scraped by performing queries using various general terms such as politics, sports, hospitals, science, cinema, economy, etc. The same query process was followed to collect reliable news articles. The configuration files utilized in this data collection process are in Appendix A.3 and Appendix A.4. The final dataset combines the data from this collection process and the unreliable and reliable data collected by Moura [61], making a total of 124 610 samples.

Table 4.1 provides an overview of the datasets utilized in the classification task. D1 corresponds to the dataset acquired through the employment of a data collection tool. D2, on the other hand, is a dataset developed by Moura [61]. By merging D1 and D2, we obtain D3, encompassing the combined data from both sources.

Table 4.1: Datasets considered for the reliability classification task.

Dataset	**Samples**	**Reliable samples**	**Unreliable samples**
Reliability Arquivo.pt dataset **(D1)**	20 644	15 854	4 790
Reliability dataset [61] **(D2)**	103 966	94 204	9 762
Large reliability dataset **(D3, D1 + D2)**	124 610	110 058	14 552

4.2.3 Exploration

The dataset used to train the model, D3, underwent a thorough exploration process to extract valuable insights for feature engineering. Three key topics were investigated to uncover these insights: named entity recognition (NER), topic modeling, and sentiment analysis. Examining these three areas aimed to understand the dataset better and identify potential features that could enhance the classification task.

The initial focus of exploration centered around named entity recognition (NER). The hypothesis put forth was that reliable news articles would contain a higher number of entity references compared to unreliable ones. The spaCy natural language processing library[1] was utilized to investigate this hypothesis. Using spaCy's entity recognition extractor, the study extracted location, organization, and person entity references from each training sample. The spaCy English named entity recognition package can identify other types, such as events, nationalities, etc. However, the Portuguese one does not contain this capability and labels all these different types with one named miscellaneous. However, due to the presence of significant noise in these miscellaneous entity references, they were ultimately disregarded in the analysis.

Figure 4.1 illustrates each news article's average number of organization, location, and person entities within the training dataset. The data indicate that reliable articles possess a significantly higher count of entity references across all categories. This contrast emphasizes the substantial disparity in entity usage between reliable and non-reliable articles.

Another interesting area of exploration is topic modeling, which involves analyzing documents to uncover clusters of related words, forming what can be considered "abstract" topics. In this study, two algorithms, Latent Dirichlet allocation (LDA) and BERTopic[2], were employed to perform topic modeling. Both algorithms are stochastic, meaning running them multiple times with the same data can yield different results. One notable advantage of BERTopic over LDA is that it eliminates the need for data pre-processing. As the name suggests, BERTopic utilizes BERT, a language model, as an embedder. This approach proves advantageous as it leverages BERT's powerful contextual representation capabilities without requiring extensive pre-processing steps. Another benefit of BERTopic is its continuous topic numbering system, eliminating the need to

[1] spacy.io

[2] github.com/MaartenGr/BERTopic

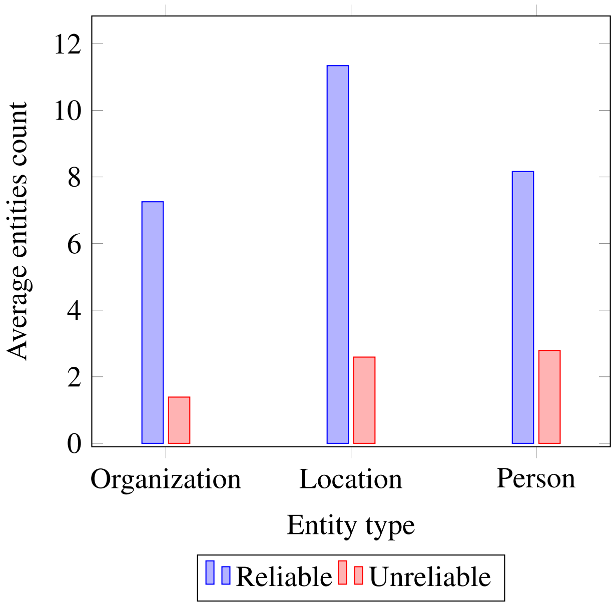


Figure 4.1: Average number of organization, location, and person entities per article in the training dataset.

specify the desired number of topics in advance. This flexibility allows the algorithm to dynamically assign topics based on the characteristics of the analyzed documents.

Table 4.2 shows the 5 topics with most documents obtained with LDA and BERTopic. Both algorithms were run to find a predefined number of 20 topics. The LDA topics were obtained by pre-processing the data by removing stop words, applying lemmatization, and stemming. Words in at least 50% of the documents and words without at least 15 appearances were removed. After that, the top 100 000 words were filtered to be considered in the corpus. The corpus used in the algorithm was based on tf-idf as the tf-idf corpus showed better results than the bag of words one. The algorithm was run for four passes. On the other hand, BERTopic required no pre-processing apart from stop words removal. BERTopic was run with a minimum topic size of 500 documents. By looking at the results in Table 4.4, it is possible to observe that the LDA results have more article distribution than the BERTopic. For instance, BERTopic merges all articles related to government, country, and president, among others, in a big topic, while the LDA distributes this topic into smaller ones.

Another hypothesis chosen to be examined is whether unreliable news articles exhibit sentiment polarity in their content. Typically, news articles maintain a neutral sentiment, although this can vary depending on the reported subject matter. Specific topics inherently carry negative emotions, such as war or crime. To analyze the sentiment, the study utilized the pysentimiento[3] [70] Python toolkit, which is a transformer-based library designed for social natural language processing tasks, including sentiment analysis and hate speech detection, among others. The toolkit supports English and three Latin languages: Portuguese, Spanish, and Italian. By employing

[3] github.com/pysentimiento/pysentimiento

Table 4.2: LDA and BERTopic topic modeling for the reliability dataset. The results show the five topics with the most documents.

(a) Five topics with most documents and their presentation created with LDA.

Topic	#documents	Representation
1	14 452	tod, ano, quer, porqu, pesso, diz, pass, muit, cas, algum
2	12 706	jog, equip, primeir, final, segund, jogador, dois, club, gol, lig
3	10 091	president, part, polít, vot, deput, parlament, govern, eleiçã, social, repúbl
4	8 665	trabalh, públic, servic, contrat, dev, ministéri, administr, govern, sindicat, segur
5	7 154	tribunal, acus, crim, process, investig, cas, públic, justic, argu, ser

(b) Five topics with most documents and their presentation created with BERTopic.

Topic	#documents	Representation
1	47 791	milhoes, governo, estado, mil, pais, estao, publico, onde, presidente, saude
2	19 030	psd, governo, presidente, ps, partido, estado, publico, republica, cds, lider
3	14 533	jogo, liga, equipa, futebol, clube, sporting, benfica, final, jogadores, fc
4	6 068	musica, festival, teatro, artistas, disco, dia, artista, album, concerto, arte
5	5 756	lisboa, porto, camara, vai, projecto, onde, empresa, publico, milhoes, area

pysentimiento, the study was able to understand the sentiment exhibited in the analyzed news articles, shedding light on any potential patterns or trends. Figure 4.2 shows the proportion of articles classified as having a neutral, negative, or positive sentiment by each label. The difference is not gigantic, but it is noticeable. Reliable articles have a proportion of 0.947 neutral articles compared to 0.806 on unreliable ones. Once again, the hypothesis put is verified as it is noticed that unreliable articles express more sentiment polarity by a smaller margin compared to reliable ones. Hence, it gives a good understanding of sentiment investigation in trustworthy news.

In this section, three key topics related to the training dataset were thoroughly examined, uncovering valuable insights that can be utilized as features in the reliability classification. Notably, the analysis revealed a potential correlation between the number of entities in a document and the expressed sentiment, potentially impacting news articles' reliability index. However, it is crucial to acknowledge that relying solely on these factors would be inadequate. Instead, they will be employed with the BERT models to augment the results and incorporate supplementary information to enhance the model's overall performance.

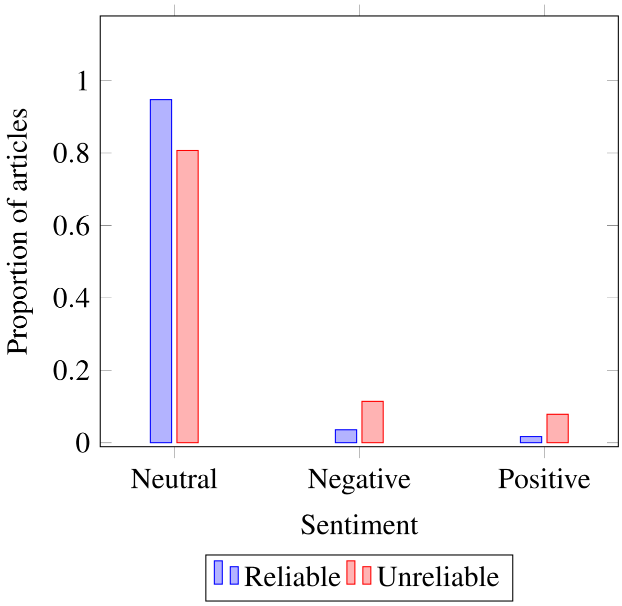


Figure 4.2: Proportion of sentiments per label in the training dataset.

4.2.4 Method

This section focuses on reliability classification in news articles and discusses the methods employed. Two strategies were utilized: a more straightforward approach and a more complex one. The simpler way is a commonly used approach for text classification tasks. On the other hand, the more complex method involves combining the standard practice with engineered features obtained through data exploration. This approach aims to enhance the classification process and provide a complete reliability assessment by incorporating additional features derived from the data exploration. This latter technique builds upon the model of the first strategy.

The dataset used in this study, D3, consists of samples labeled as reliable or unreliable. However, upon examining Table 4.1, it becomes apparent that the dataset is highly unbalanced. The majority class, representing reliable news, accounts for 88% of the dataset. It is crucial to consider this class imbalance when evaluating the performance metrics of the model. To address this issue, two metrics were considered: the macro-averaged F1 score and accuracy. In single-label classification problems, accuracy is equivalent to a micro-averaged F1 score. However, relying solely on accuracy can be misleading in the case of imbalanced datasets like this one. For instance, if the model predicts all samples as belonging to the majority class, it could achieve a high accuracy score of 88% (assuming the distribution of classes in the training set is the same in the validation set). This outcome does not accurately assess the model's actual performance. To overcome this limitation, the F1 macro score was utilized. By macro-averaging the F1 scores per class, this metric offers a more reliable evaluation. It treats all categories equally, regardless of their support values. This approach ensures that each type contributes similarly to the overall assessment, thereby addressing the issue of class imbalance.

Both methods used the same dataset split of 80%-20% for training and validation, using a stratified random split. The samples were pre-processed before feeding the model with a light pre-processing that removed multiple spaces, links, and hashtags. The loss function used to train the models was the Binary Cross-Entropy loss considering the class weights. The BERT models used the maximum input length, which is 512. The optimizer was AdamW, an exponential scheduler for BERTimbau and XLM-RoBERTa, and a linear scheduler with a warmup in the Albertina-ptpt large. The last was motivated by the small batch size employed in Albertina-ptpt large. When using a small batch size, the gradient estimates can be noisy and less representative of the underlying data distribution. This can lead to high variance in the training process, resulting in unstable updates and slower convergence. Warmup helps by initially utilizing a lower learning rate, allowing the model to make more stable updates based on a broader range of samples. As the learning rate gradually increases, the model becomes more confident in its parameter updates and can better adapt to the data distribution.

Method 1 Common text classification method

This strategy involves adding a linear layer on top of the last hidden layer of BERT to perform classification. The architecture typically includes transitioning from the context encoding of the last hidden layer of BERT to a classifier layer with a single neuron. A dropout layer is incorporated between the BERT encodings and the classification layer to enhance the model's robustness and prevent overfitting. This dropout layer randomly sets a fraction (0.15) of the activations to zero during training, helping to regularize the model and reduce the reliance on specific hidden units. Table 4.3 shows the BERTimbau, XLM-RoBERTa, and Albertina-ptpt configuration for this method.

Table 4.3: Reliability models' configuration using method 1.

Model	Learning rate	Batch size	Steps/epoch	Scheduler
BERTimbau$_{\text{BASE}}$ XLM-RoBERTa$_{\text{BASE}}$ Albertina-ptpt$_{\text{BASE}}$	2×10^{-5}	28	3 560	Exponential $\gamma = 0.99985$
BERTimbau$_{\text{LARGE}}$		20	4 984	Exponential $\gamma = 0.99990$
XLM-RoBERTa$_{\text{LARGE}}$		16	6 230	Exponential $\gamma = 0.99995$
Albertina-ptpt$_{\text{LARGE}}$		6	16 614	Linear with 1 500 warmup steps

Method 2 Feature concatenation

The second method builds upon the first one. This method uses an approach known as feature concatenation, which consists of appending the extra features to the last hidden layer of BERT. The features used were selected by analyzing the exploration phase of the training dataset.

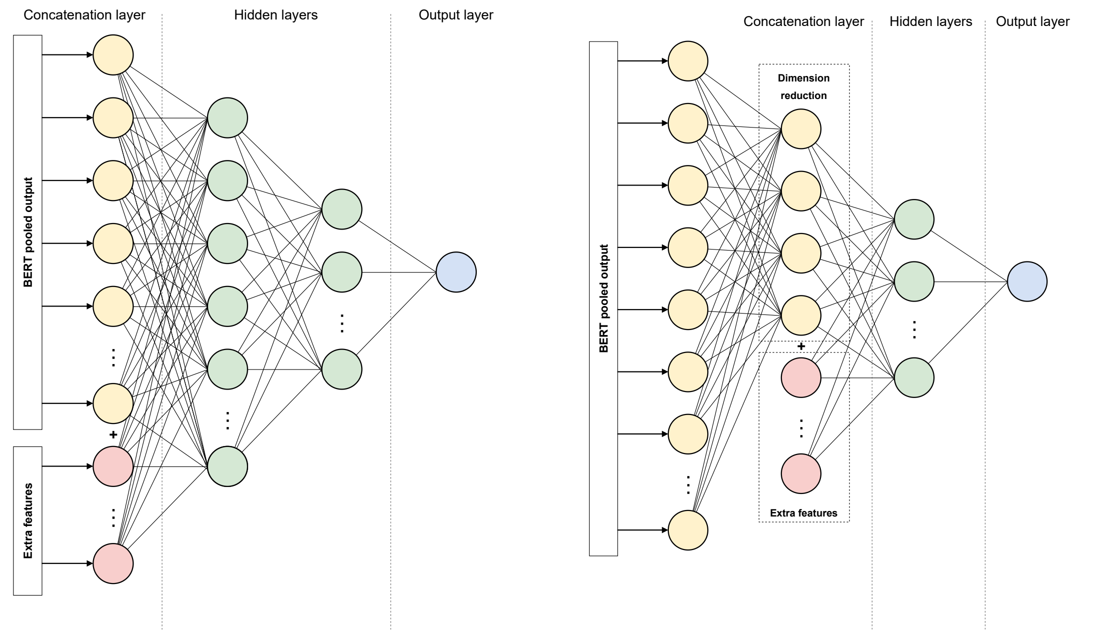
They are the article's topic, the sentiment, the probability of each sentiment, and the entity, organization, and person entities count. Table 4.4 shows the extra features used in this approach. Before feeding the data to the model, the features were pre-processed. The categorical features were one-hot encoded, and the numerical features were normalized using the min-max feature scaling.

Table 4.4: Extra features, their description, and type, used in the method feature concatenation.

Feature	Description	Type
Sentiment	The sentiment type of the article	Categorical
Neutral sentiment probability	The neutral sentiment probability of the article	Numerical
Negative sentiment probability	The negative sentiment probability of the article	Numerical
Positive sentiment probability	The positive sentiment probability of the article	Numerical
Organization entities count	The number of organizations entities of the article	Numerical
Location entities count	The number of location entities of the article	Numerical
Person entities count	The number of person entities of the article	Numerical
Total entities count	The total number of entities of the article	Numerical
Topic	The topic of the article	Categorical

The neural network employed in this context is more intricate than its predecessor. Typically, the standard procedure involves concatenating the features and utilizing a multi-layer perceptron for further processing. However, this method was tested using two techniques: (1) concatenation followed by a multi-layer perceptron, and (2) applying a linear layer on context encodings to reduce the number of neurons, followed by concatenation and a multi-layer perceptron. Figure 4.3 illustrates the architecture employed in both models. The architectures are simplified for explaining purposes. Their starting point is on BERT pooled output. The number of hidden layers and neurons represented is exclusively arbitrary. The description of the architecture used in the models, such as the number of hidden layers, neurons, and dropouts, among others, is described in the next section.

The last hidden layer of BERT models often has high dimensions, such as 768 or 1 024. Introducing a linear layer after the context encodings helps reduce the size of the encodings. This experimental approach aimed to verify whether the model could utilize the additional features more efficiently when employing dimension reduction. This was particularly important because, without dimension reduction, the context embeddings' dimension would be significantly smaller than the extra features' dimension. Several experimental results have demonstrated that the reduction approach consistently had slightly better results than the non-reduction method. This reduction technique balances the proportion of neurons from the BERT model and the additional features present in the concatenation layer. As a result, the model gains a better understanding of



(a) Feature concatenation followed by multi-layer perceptron architecture.

(b) BERT pooled output reduction then feature concatenation followed by multi-layer perceptron architecture.

Figure 4.3: Different approaches, (1) and (2), to feature concatenation .

the weights associated with the extra features than when the reduction technique is not employed. Hence, the dimension reduction approach was used in the further trained models.

The initial stage of training this method involves acquiring the pre-trained weights and biases from the top-performing model of method 1. Subsequently, the parameters of the BERT model are set to a frozen state, and the sole focus lies on training the parameters of the multi-layer perceptron. BERT-based models commonly employ a learning rate within the range of 2×10^{-5} to 5×10^{-5}. However, this learning rate is relatively low for the multi-layer perceptron. To speed up the training process, the BERT model is frozen, ensuring that only the parameters of the multi-layer perceptron undergo training. Subsequently, the next step occurs when the model reaches a point where further improvements in validation loss become challenging. All model parameters are unfrozen at this stage, allowing them to become trainable again. The model is then retrained, employing a significantly lower learning rate of 5×10^{-6}.

For reference, the architecture of the multi-layer perceptron employed in each model can be found in Table 4.5. Between every hidden layer is a batch normalization layer and a dropout layer.

4.2.5 Evaluation

This section explores the results of the method of the previous section using the BERTimbau, XLM-RoBERTa, and Albertina-ptpt models. The BERT model was initially frozen to incorporate the extra features and fine-tune the model using method 2. During this phase, all models were

Table 4.5: Architectures to fine-tune the models using feature concatenation.

			Hidden layers		
Model	**Pooled output size**	**Reduction size**	**Number**	**Size**	**Dropout**
BERTimbau$_{\text{BASE}}$	768	128	2	64, 32	0.15
XLM-RoBERTa$_{\text{BASE}}$					
Albertina-ptpt$_{\text{BASE}}$					
BERTimbau$_{\text{LARGE}}$	1 024	192	2	96, 48	0.15
XLM-RoBERTa$_{\text{LARGE}}$					
Albertina-ptpt$_{\text{LARGE}}$	1 536	256	3	128,64,32	

trained with a batch size of 48 and a constant learning rate 0.01. Once all parameters were made trainable again, the batch size for each model remained consistent with method 1, as outlined in Table 4.3. Furthermore, a constant learning rate of 5×10^{-6} was applied to all models.

Table 4.6 shows the best results of the architectures defined above.

Table 4.6: Results obtained for the reliability classification using BERTimbau, XLM-RoBERTa, and Albertina-ptpt.

Model	**Accuracy**	**F1$_{\text{macro}}$**	**F1$_{\text{weighted}}$**
BERTimbau$_{\text{BASE}}$	0.9896	0.9719	0.9872
XLM-RoBERTa$_{\text{BASE}}$	0.9897	0.9717	0.9880
Albertina-ptpt$_{\text{BASE}}$	0.9851	0.9601	0.9817
BERTimbau$_{\text{LARGE}}$	**0.9919**	**0.9811**	**0.9905**
XLM-RoBERTa$_{\text{LARGE}}$	0.9910	0.9791	0.9890
Albertina-ptpt$_{\text{LARGE}}$	0.9899	0.9731	0.9876

These meticulous and incremental steps in this axis yield impressive results, with an accuracy of 0.9919 and an F_1 macro score of 0.9811. These outcomes surpass the performance exhibited by Moura using a BERT model alongside computational forensic features in a superset of his training dataset, showcasing a substantial enhancement in the average F_1 macro score from 0.96 to 0.9811.

4.3 Political Stance

In this work, political stance refers to an individual's, or in this case, a news article, position, or viewpoint on various political issues, ideologies, or beliefs. It represents a perspective on governance, social issues, economics, international relations, and more. This section focuses on analyzing and developing a method for detecting political stances in Portuguese news. The process begins with an examination of existing research on political stance detection. Following that, the datasets employed in the detection approach are defined and explained. Subsequently, the methodology used to construct the model is illustrated, highlighting using a BERT-based model as a critical component, as the previous axis. Finally, the algorithm's performance is evaluated using pertinent metrics to assess its effectiveness.

4.3.1 Related Work

No previous efforts have focused on the automated detection of political stances in Portuguese news. While relevant studies within the political sciences or related areas exist [71], they do not specifically address the automated detection aspect. However, when expanding the scope beyond Portuguese, valuable research is available, particularly concerning the utilization of BERT-based models [54, 55], which aligns with the approach taken in the previous phases of this work.

Graça [71] conducted an extensive study on Portuguese political bias, focusing on the period from 2009 to 2015. The study employed manual approaches, as described in chapter 2, to detect political media bias. Additionally, the tone of the news articles was carefully evaluated as part of the analysis. Through this meticulous examination, Graça successfully identified patterns indicating that certain media outlets exhibited preferences for specific political orientations. This discovery sheds light on the political bias study for Portuguese news, revealing a real problem.

Continuing within the context of Portugal, Aparicio et al. [72] conducted a noteworthy study investigating the utilization of emotions in the communication strategies of Portuguese political parties on their official Twitter accounts. Their research explicitly centered on the period coinciding with the COVID-19 pandemic. The study revealed distinct variations in the emotional content expressed within the communication of different political parties. However, it is essential to note that the results showed striking similarities across all parties, likely due to the single-domain nature of the topic under investigation, namely the COVID-19 pandemic. This indicates that the prevailing circumstances during the specified period played a significant role in shaping the emotional tones adopted by the political parties across the board.

Beyond the Portuguese content, substantial research has focused on identifying political stances in English news articles. Online bias indexing platforms such as Media Bias Fact Check or AllSides have emerged as valuable resources, facilitating the collecting and categorizing news articles based on their political orientation. This approach proves instrumental in generating political stance datasets. Numerous studies have successfully employed this methodology, gathering datasets and training BERT-based models to detect political bias [54, 55], yielding promising outcomes. However, the current main challenge within the political landscape lies in the scarcity of datasets. The absence of a political bias indexer for Portuguese news presents a barrier, hindering the pursuit of this research.

4.3.2 Datasets

As previously mentioned, the absence of Portuguese political stance datasets poses a significant challenge to this research. Attempting to translate an English dataset is not a viable solution, as a foreign country's political landscape and context can vary significantly from that of Portugal. However, alternative resources could be leveraged to overcome this obstacle. One potential source of valuable data lies in the parliament interventions of political parties. These interventions can offer insights into different parties' political stances and positions, providing a foundation for constructing a dataset specific to the Portuguese context. While not without limitations, this approach presents an opportunity to gather relevant information and address the lack of available datasets in the political panorama.

A dataset of Portuguese parliamentary minutes was compiled containing the interventions made by deputies during legislative sessions. Extracting information from these minutes established a collection of interventions associated with respective political parties. The dataset curated comprises 5 713 parliament interventions originating from deputies belonging to ten different political parties. Recognizing the initial scarcity of the data, efforts were made to augment it using OpenAI's powerful `gpt-3.5-turbo` model. The augmentation process involved summarizing the political interventions. Appendix C provides an example of the API requests and corresponding responses used in the augmentation process. The prompt asked to sum up an intervention, capturing its key points. Due to usage limits and API costs, the augmentation process was only partially completed. Nevertheless, this augmentation yielded a final dataset comprising 8 881 samples, expanding the original dataset's volume and diversity.

To assign a political orientation label to each intervention, the Wikipedia page of Portuguese political parties was utilized as a valuable resource[4]. While there is no official designation for the political orientation of each party, the Wikipedia pages serve as an informative platform that is regularly accessed and edited by numerous individuals daily. This collective effort ensures an exhaustive categorization of the political orientations of Portuguese parties, as it reflects the consensus agreed upon by a group of contributors. Thus, the Wikipedia pages were believed to be reliable for determining the political orientations associated with each party and were utilized as a reference for labeling the interventions accordingly.

With Wikipedia information, two types of labels were defined: strict and a relaxed one. The strict label has five unique values – far-left, left-wing, center, right-wing, and far-right. In comparison, the relaxed one has three unique values – lean left, center, and lean right. Table 4.7 shows each Portuguese party's strict and relaxed labels. The strict label was directly obtained from the party's label. The assigned label was the most away from the center when the party had two labels. The relaxed label is acquired from the strict label, eliminating the far-left and far-right instances, turning them into lean-left and lean-right, respectively.

In conclusion, the distribution of samples in the dataset is presented in Table 4.8. The data

[4] pt.wikipedia.org/wiki/Lista_de_partidos_pol%C3%ADticos_em_Portugal

Table 4.7: Portuguese political parties and their respective strict and relaxed labels derived from Wikipedia's party label.

Party	Wikipedia label	Strict label	Relaxed label
PCP	Left to far-left	Far-left	Lean-left
BE	Left-wing	Left-wing	Lean-left
PEV	Left-wing	Left-wing	Lean-left
L	Center-left to left-wing	Left-wing	Lean-left
PS	Center-left	Center	Center
PAN	Center-left	Center	Center
PSD	Center-right to right-wing	Right-wing	Lean-right
IL	Right-wing	Right-wing	Lean-right
CDS-PP	Right-wing	Right-wing	Lean-right
CH	Far-right	Far-right	Lean-right

reveals that the dataset is not evenly balanced regarding the strict label, with only 100 samples representing the far-right category. However, when considering the relaxed label, which allows for a broader classification approach, the dataset exhibits a more uniform distribution across the different labels. This dataset is not ideal as it is based on a domain different from the domain of news articles. However, considering the absence of alternatives, it is a great foundation for this work.

Table 4.8: Strict label and relaxed label distributions in the political dataset.

Strict label

Far-left	Left-wing	Center	Right-wing	Far-right
1550 (17.5%)	1834 (20.7%)	1939 (21.8%)	3458 (38.9%)	100 (1.1%)

Relaxed label

Lean-left	Center	Lean-right
3384 (38.1%)	1939 (21.8%)	3558 (40.1%)

4.3.3 Exploration

The dataset underwent an exploration process to extract meaningful insights, explicitly focusing on topic modeling. The decision to narrow the analysis to this aspect was based on the lack of substantial evidence in previous works regarding the effectiveness of alternative studies similar to those conducted in the earlier axes.

Similarly to the reliability axis, the exploration phase also used both Latent Dirichlet Allocation (LDA) and BERTopic techniques to extract topics from the political documents. The consideration of topics in this exploration was motivated by the observation that certain political parties concentrate their efforts within specific domains. By gathering the extracted topics from the samples and subjecting the data to analysis, valuable insights can be gained regarding the usefulness of this information for the model. This examination comprehends the relevance and potential impact of incorporating topic-based features into the model's decision-making process.

Table 4.9 presents the top 5 topics identified using LDA and BERTopic, both run with a predefined number of 10 topics. The pre-processing applied to both models followed the same methodology used in the reliability axis. Figure 4.4 illustrates the distribution of sample proportions per topic obtained from the LDA tf-idf algorithm. The results obtained from LDA and BERTopic exhibited notable similarities, with the balance of articles per topic being nearly identical for each label. This suggests that political parties generally address the same topics. Given that the dataset comprises Portuguese parliamentary interventions during parliamentary sessions, it is expected that the topics covered by the parties are reasonably consistent, as they contribute to discussions on these topics in similar efforts.

The topic modeling analysis proved ineffective in this context, as the distribution of samples per topic did not significantly differ across labels. Therefore, this axis did not incorporate the topic modeling feature and instead followed a standard approach.

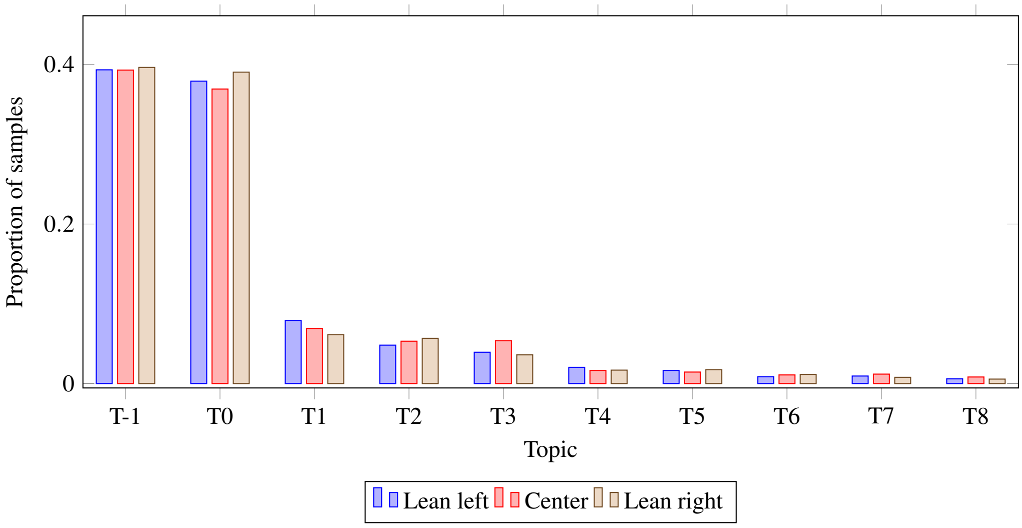


Figure 4.4: Proportion of sample' topics per label in the political stance dataset using LDA with tf-idf.

Table 4.9: LDA and BERTopic topic modeling for the political stance dataset. The results show the five topics with the most documents.

(a) Five topics with most documents and their presentation created with LDA.

Topic	#documents	Representation
1	3 503	materia, forma, portanto, empresas, seguranca, so, dizer, regime, questao, trabalho
2	3 388	trabalho, trabalhadores, materia, portugueses, so, seguranca, dizer, nacional, forma, debate
3	620	madeira, regiao, assembleia, autonoma, regioes, verdes, autonomas, legislativa, florestal, regional
4	468	ensino, escola, escolas, educacao, alunos, superior, autonomia, professores, gestao, escolar
5	365	saude, cuidados, medicamentos, servico, nacional, farmacias, sns, hospitais, acesso, taxas

(b) Five topics with most documents and their presentation created with BERTopic.

Topic	#documents	Representation
1	5 547	trabalh, públic, empres, fiscal, quer, diz, administr, secretári, matér, segur
2	1 078	escol, ensin, educ, autor, alun, criminal, text, projet, crim, alto
3	943	autor, text, med, públic, direit, fiscal, águ, projet, aument, florestal
4	566	saúd, trabalh, servic, públic, autor, projet, cuid, text, moder, nacional
5	241	regiã, madeir, autónom, autor, text, desport, ped, projet, açor, regional

4.3.4 Method

This section focuses on classifying political texts, building upon the groundwork in earlier sections. As discussed previously, two distinct label types were defined, necessitating the adoption of two strategies- one for each label type. As previously stated, it is essential to note that the dataset used in this study is not explicitly comprised of news articles.

In this axis, the dataset used is unbalanced, as previously discussed. Therefore, the classification metrics of primary interest are consistent with those utilized in the previous axes, namely the macro-averaged F1 score, and accuracy. The dataset split consisted of a stratified random split of 80%-20% for training and validation, respectively. This split was done on the initial non-augmented dataset. After the split, the augmented samples with `gpt-3.5-turbo` were concatenated to the training samples, ensuring that the augmented samples were only used in training. The rest of the configuration follows the identical steps of the previous axis, having a 512 maximum input length, light text pre-processing, AdamW as the optimizer, and a learning rate of 2×10^{-5}. However, the exponential scheduler γ value and the linear warmup steps were modified to match the dataset size.

Table 4.10 shows the models' configuration. However, the loss function used in this axis is a custom loss function built together with the cross-entropy loss.

Table 4.10: Political models' configuration.

Model	**Learning rate**	**Batch size**	**Steps/epoch**	**Scheduler**
BERTimbau$_{\text{BASE}}$	2×10^{-5}	28	276	Exponential $\gamma = 0.9990$
XLM-RoBERTa$_{\text{BASE}}$				
Albertina-ptpt$_{\text{BASE}}$				
BERTimbau$_{\text{LARGE}}$		20	386	Exponential $\gamma = 0.9993$
XLM-RoBERTa$_{\text{LARGE}}$		16	483	Exponential $\gamma = 0.9994$
Albertina-ptpt$_{\text{LARGE}}$		6	1289	Linear with 150 warmup steps

The labels themselves have an intrinsic proximity property, *e.g.*, considering the strict label, the far-left is closer to the left wing than the center or the right-wing labels. The loss function was modified to incorporate an extra cost to the cross-entropy loss value to account for this property. Equation (4.1) shows the cross-entropy loss formula for a sample considering class weights. In this case, the cross-entropy formula is equivalent to combining the log softmax and the negative log-likelihood loss. Equation (4.2) shows the cross-entropy loss formula for a batch of samples.

$$l(s) = -w_{y_s} \log \frac{\exp(x_{s,y_s})}{\sum_{i=1}^{C} \exp(x_{s,i})} \tag{4.1}$$

x_s are the logits for the sample s, y_s is the target, w is the weight and C the number of classes.

$$\ell = \sum_{s=1}^{S} \frac{1}{\sum_{s=1}^{S} w_{y_s}} \cdot l(s) \tag{4.2}$$

S is the number of samples of a batch, l_s is the sample s cross-entropy loss.

The extra cost was implemented to assign a lower value to predictions that center their probabilities in the target or proximity classes. As a result, higher values were assigned to forecasts that failed to guess the target value accurately and instead gave higher probabilities to labels that were further away from the desired class. This approach encourages the model to consider the notion of class proximity, giving higher chances to related classes and avoiding, for instance, having high probabilities for the two classes farther apart. Equation (4.3) shows the extra cost for a sample s. A sample's extra cost is obtained by the sum of multiplying the softmax probabilities of each class with the absolute unit distance, calculated with the distance divided by the max distance. Table 4.11 shows the distances between every pair of labels.

$$l_{\text{extra-cost}}(s) = -\sum_{c=1}^{C} \left[\left| \frac{\mathrm{d}(c, y_s)}{max\_distance} \right| \cdot \frac{\exp(x_{s,c})}{\sum_{i=1}^{C} \exp(x_{s,i})} \right] \tag{4.3}$$

x_s are the logits for the sample s, y_s is the target, and C the number of classes. The $\mathrm{d}(c, y_s)$ calculates the class distance between c, an arbitrary class, and the target class, y_s. $max\_distance$ is the maximum distance between a pair of classes.

$$\ell_{\text{extra-cost}} = \frac{1}{S} \sum_{s=1}^{S} l_{\text{extra-cost}}(s) \tag{4.4}$$

S is the number of samples of a batch, le_s is the sample s extra cost.

The value of $\ell_{\text{extra-cost}}$ is always between 0 and 1, as it is the average of $l_{\text{extra-cost}}(s)$, which is also between 0 and 1. The custom loss is then implemented by multiplying the cross-entropy loss value by $1 + \ell_{\text{extra-cost}}$. Equation (4.5) shows the formula for the custom loss.

$$loss = \ell(1 + \phi \ell_{\text{extra-cost}}) \tag{4.5}$$

ϕ is a multiplicative factor to regulate the weight of the extra cost. The default value and the one used in the training was $\phi = 1$.

Table 4.11: Absolute distances between the labels on the political dataset.

Label/distance	**Far-left**	**Lean-left** **Left-wing**	**Center**	**Lean-right** **Right-wing**	**Far-right**
Far-left	0	1	2	3	4
Lean-left/Left-wing	1	0	1	2	3
Center	2	1	0	1	2
Lean-right/Right-wing	3	2	1	0	1
Far-right	4	3	2	1	0

4.3.5 Evaluation

This section explores the results obtained by training the models previously described with the strict and relaxed label. Table 4.12 and Table 4.13 show the results for the strict and relaxed labels, respectively. The results are satisfactory, with the BERTimbau$_{\text{LARGE}}$ model achieving the best results in both labels. The F1 macro score is notable because the dataset is highly unbalanced, particularly when considering the task with the strict label and considering the far-right class, which has only 1.1% of the data samples. As expected, the results are particularly better for the relaxed version of the dataset, which reduces the number of existing labels to classify.

The performance of large models (those with more than 300 million parameters) in this dataset has been subpar. This can be attributed to the fact that larger models generally require more data to generalize effectively. While the results obtained so far are satisfactory, they are not particularly impressive, primarily because the dataset does not adequately represent the real-world domain of news. To significantly enhance the performance of the task, a crucial improvement would be to gather a more suitable dataset that aligns closely with the actual characteristics and nuances of news articles. By doing so, the model's performance can be expected to improve substantially.

Table 4.12: Results obtained for the political stance classification with the strict labels using BERTimbau, XLM-RoBERTa, and Albertina-ptpt.

Model	**Accuracy**	**F1$_{macro}$**	**F1$_{weighted}$**
BERTimbau$_{BASE}$	0.7419	0.6539	0.7312
XLM-RoBERTa$_{BASE}$	0.6884	0.6015	0.6815
Albertina-ptpt$_{BASE}$	0.6450	0.5690	0.6457
BERTimbau$_{LARGE}$	**0.7777**	**0.7091**	**0.7709**
XLM-RoBERTa$_{LARGE}$	0.6911	0.6000	0.6868
Albertina-ptpt$_{LARGE}$	0.6766	0.4925	0.6766

Table 4.13: Results obtained for the political stance classification with the relaxed labels using BERTimbau, XLM-RoBERTa, and Albertina-ptpt.

Model	**Accuracy**	**F1$_{macro}$**	**F1$_{weighted}$**
BERTimbau$_{BASE}$	0.7760	0.7183	0.7721
XLM-RoBERTa$_{BASE}$	0.6823	0.6149	0.6858
Albertina-ptpt$_{BASE}$	0.6369	0.5551	0.6253
BERTimbau$_{LARGE}$	**0.8022**	**0.7461**	**0.8052**
XLM-RoBERTa$_{LARGE}$	0.6994	0.6352	0.6978
Albertina-ptpt$_{LARGE}$	0.6815	0.5142	0.6815

4.4 Objectivity

As previously defined in Chapter 3, the concept of objectivity revolves around the deliberate exclusion of subjective opinions and biased terminology. This section aims to enhance our understanding of objectivity in news texts by presenting a methodical approach to its classification. The initial step involves analyzing existing research and studies on this subject matter. Subsequently, a description of the datasets employed in developing the model is provided. Following that, an elaborate explanation of the model training process is presented, which is thereafter followed by an evaluation of its performance.

4.4.1 Related Work

The classification of objectivity or subjectivity, commonly referred to as sentence-level subjectivity detection in research, is a well-established subject in natural language processing. While BERT-based models have gained prominence recently, it is worth noting that significant progress has been made in this area using alternative methodologies before their rise. Such methodologies include rule-based approaches [73] or the adaptation of LDA (Latent Dirichlet Allocation) [74] to develop a weakly-supervised generative model for subjectivity detection.

Furthermore, with the advent of BERT-based models, a significant shift occurred in sentence-level subjectivity classification [20, 21]. BERT facilitated pre-training and fine-tuning techniques. Researchers employed BERT as a feature extractor or fine-tuned the model on specific subjectivity detection tasks, attaining state-of-the-art performance. The contextual embeddings captured by BERT enabled a more comprehensive understanding of sentence semantics and improved the accuracy of subjectivity classification.

4.4.2 Datasets

In 2021, Spinde et al. [38] conducted a comprehensive research study on media bias, which involved the creation of an expert annotated dataset known as BABE (Bias Annotations By Experts). This dataset comprised 3700 news articles, with annotations provided at the sentence level to identify opinions expressed within the text. As detailed in Chapter 3, the annotation process involved two groups of annotators: Group 1 and Group 2. Given that multiple annotators evaluated each sentence, the labels were assigned based on a majority vote. The labels used for annotation included “opinionated”, “factual”, “mixed”, and “no agreement”. The distribution of dataset labels within the BABE dataset is presented in Table 4.14.

Table 4.14: Label and relaxed label distributions in the objectivity dataset.

Opionated	**Factual**	**Mixed**	**No agreement**
1628 (44%)	951 (25.7%)	874 (23.6%)	247 (6.7%)

To prepare the data for model training, sentences labeled “no agreement” were excluded from the dataset, resulting in a final sample size of 3453 instances. These samples, which exhibited clear consensus among the annotators, were retained to ensure the integrity and reliability of the dataset for subsequent model analysis.

4.4.3 Method

This section describes the method used to train the model. It builds upon the groundwork done in the previous axis modifying the configuration of reliability’s method 1 to classify the sentences.

The dataset was divided using a stratified random split, allocating 80% of the data for training and 20% for validation. The evaluation metrics considered for assessing model performance were the macro-averaged F1 score and accuracy. For this particular axis, the configuration is almost identical to the previous one, which focused on political stance. However, some modifications were made to the scheduler parameters and the loss function. The scheduler parameters were adjusted accordingly to adapt to the dataset size and the number of steps per epoch. These changes can be observed in Table 4.16.

Furthermore, the custom loss defined in equation (4.5) was also applied is $\phi = 1$. This decision was influenced by the inherent characteristic of the labels, which exhibit a sense of proximity between them. The custom loss function was designed to account for this proximity. For example, sentences labeled as “opinionated” are closer to sentences labeled as “mixed” than sentences labeled as “factual”. To illustrate the degree of proximity between the labels, Table 4.15 shows the distance between the labels. The maximum distance observed in the Table is 2, indicating that the furthest two labels can be from each other in proximity.

Table 4.15: Absolute distances between the labels on the objectivity dataset.

Label/distance	**Opinionated**	**Mixed**	**Factual**
Opinionated	0	1	2
Mixed	1	0	1
Factual	2	1	0

4.4.4 Evaluation

The model was trained using the same models as before, following the configuration described earlier. The results obtained for the validation set are presented in Table 4.17, showing the accuracy, F1 macro, and F1 weighted metrics.

Table 4.16: Objectivity models' configuration.

Model	Learning rate	Batch size	Steps/epoch	Scheduler
BERTimbau$_{\text{BASE}}$ XLM-RoBERTa$_{\text{BASE}}$ Albertina-ptpt$_{\text{BASE}}$	2×10^{-5}	28	98	Exponential $\gamma = 0.9980$
BERTimbau$_{\text{LARGE}}$		20	138	Exponential $\gamma = 0.9985$
XLM-RoBERTa$_{\text{LARGE}}$		16	172	Exponential $\gamma = 0.9990$
Albertina-ptpt$_{\text{LARGE}}$		6	460	Linear with 50 warmup steps

Table 4.17: Results obtained for the objectivity classification using BERTimbau, XLM-RoBERTa, and Albertina-ptpt.

Model	Accuracy	F1$_{\text{macro}}$	F1$_{\text{weighted}}$
BERTimbau$_{\text{BASE}}$	0.8070	0.7524	0.8040
XLM-RoBERTa$_{\text{BASE}}$	0.7980	0.7494	0.7988
Albertina-ptpt$_{\text{BASE}}$	0.7910	0.7227	0.7830
BERTimbau$_{\text{LARGE}}$	**0.8260**	**0.7810**	**0.8275**
XLM-RoBERTa$_{\text{LARGE}}$	0.8032	0.7350	0.8041
Albertina-ptpt$_{\text{LARGE}}$	0.7739	0.2580	0.7739

Despite the inherent limitations of the dataset, including its limited size and the need for translation, the obtained results are satisfactory. Among the evaluated models, BERTimbau$_{\text{LARGE}}$ achieved the best performance, with an accuracy score of 0.8260 and an F1 macro of 0.7810. These results suggest that the model can perform well overall and for each class, as indicated by the good F1 macro score. It is worth noting that the dataset was slightly unbalanced, yet the model demonstrated the ability to identify the minority classes compared to the majority classes effectively.

On the other hand, Albertina-ptpt$_{\text{LARGE}}$ exhibited the lowest performance, despite performing better than BERTimbau in other tasks [64]. This can be attributed to Albertina-ptpt$_{\text{LARGE}}$ being a large transformer model with 900 million parameters. However, due to the limited size of the dataset, the available data are insufficient for the model to generalize and perform well effectively. The Albertina model was also trained in less data than the BERTimbau model.

The results are satisfactory, but there is potential for further improvement. One crucial aspect is the expansion of the dataset to include a more significant amount of data. By incorporating more data, conducting a more comprehensive dataset analysis would be possible, providing valuable insights for model development. Additionally, increasing the dataset size would benefit the more powerful models in the list, as the current amount of data appears insufficient for optimal performance. Expanding the dataset would offer these models more abundant and diverse training examples, enabling them to leverage their capabilities better.

4.5 Readability

Readability refers to the degree of comprehensibility or understandability of a text. This section analyzes different indices used to measure readability and provides an overview of their functioning. It is necessary to mention that these indices are formula-based metrics designed to assess the understandability of a text. Thus, this section does not explicitly evaluate readability but focuses on clarifying the mechanisms underlying these indices.

4.5.1 Readability Indices and Formulas

The work on readability indices encloses various studies and research conducted to explore and develop metrics that quantitatively measure the readability of texts. Numerous readability indices have been proposed, each with its approach and formula. These indices aim to provide objective measures of text complexity and comprehensibility.

Recent research has also explored machine learning techniques to develop more sophisticated readability models. These models often utilize linguistic features, semantic analysis, and contextual information to predict text readability accurately. The related work on readability indices spans several decades and encompasses traditional formula-based approaches and more recent advancements in machine learning. These indices provide valuable tools for assessing text complexity and aiding in creating accessible and comprehensible written content.

In this particular study, the focus was not on utilizing machine learning approaches for assessing readability. Instead, the emphasis was placed on exploring traditional methods. This choice was because while machine learning techniques have been extensively employed in the previous axes, a more straightforward yet precise and accurate approach seemed more suitable for the current axis. Unlike computer-intensive machine learning algorithms, the formulas used for readability calculations are less computationally demanding. This characteristic enables the analysis of multiple readability indices, allowing news consumers to choose the index that best suits their preferences. Further exploration of this concept is presented in Chapter 5.

The indices and formulas discussed were initially designed for the English language. However, it is worth noting that a Portuguese study conducted by Moreno et al. [75] successfully adapted some of these formulas to suit the nuances of the Portuguese language. The adaptation process involved regression analysis, as all subsequent formulas exhibit a linear dependence on two variables. This adjustment allowed for a more accurate assessment of readability in Portuguese texts, providing valuable insights into the readability levels for a broader range of readers.

One significant challenge in readability formulas, particularly those that rely on syllable counting, is accurately determining the number of syllables in a word. Even today, this remains a difficult task, and there is no algorithm capable of precisely calculating the number of syllables in Portuguese words. However, the Center for General and Applied Linguistics Studies (CELGA-ILTEC) provides in the Portal of the Portuguese Language[5] a list of 181 990 words along with their corresponding syllable counts. By using this resource, it is possible to get the number of syllables of words accurately.

Nevertheless, when the word is not available in the Portal of the Portuguese Language list, an algorithm is employed. The algorithm is based on the Moreno et al. [75] work. The algorithm can correctly calculate the number of syllables of 82.66% of the previous corpus of 181 900 words. When the algorithm fails to identify the number of syllables correctly, the mean difference to the correct number of syllables is 1.07 with a standard deviation of 0.26. The algorithm begins by storing the vowels, diphthongs, and triphthongs that exist in the Portuguese language. Subsequently, it iterates through the word's letters, incrementing a counter whenever a vowel is encountered. Simultaneously, the algorithm decrements the count for semivowels within diphthongs and triphthongs. In the case of diphthongs, the algorithm reduces the counter by one if a diphthong is identified and the preceding letter is not a vowel. Similarly, for triphthongs, the algorithm decreases the counter by one upon detecting a triphthong. Algorithm 1 shows the pseudocode for the algorithm to estimate the number of syllables of a word.

Some of the following formulas provide a result interpreted by the U.S. grade level, essentially identical to the Portuguese grade level. Essentially the scores will represent the grade required by the reader to be able to understand the text swiftly. Scores from 1 to 12 are for regular education grade levels. The ones above 13 are for college students or graduates.

The following section delves into the leading indices and formulas significantly impacting readability estimation.

4.5.1.1 Flesch Reading Ease and Flesch-Kincaid Grade Level

A notable contribution to the field of readability assessment is the Flesch reading ease formula, introduced by Rudolf Flesch in 1948 [76]. This formula utilizes the average number of syllables per word and the average number of words per sentence to derive a readability score. The resulting score is normalized on a scale of 0 to 100, where 0 represents minimal readability, and 100 corresponds to maximum readability. Later, in 1975, J. Peter Kincaid modified the Flesch Reading Ease formula to fit the U.S. grade level, making it for people to understand.

Equation (4.6) is the Flesch reading ease formula, and equation (4.7) is the Flesch-Kincaid grade level.

[5] www.portaldalinguaportuguesa.org

Algorithm 1 Algorithm to estimate the number of syllables in Portuguese words.

```
$vowels \leftarrow [a, ã, â, á, à, e, é, ê, i, í, o, ô, õ, ó, u, ú]$
$dipthtongs \leftarrow [ãe, ai, ão, au, ei, éi, eu, éu, ia, ie, io, iu, õe, oi, ói, ou, ua, ue, uê, ui]$
$triphthongs \leftarrow [uai, uei, uão, uõe, uiu, uou, uem, uam, ual, uai]$

procedure CALCULATE_SYLLABLES($word$)
    $syllables \leftarrow 0$
    $word\_length \leftarrow length(word)$

    for $idx$ in $0 \ldots word\_length$ do
        if $word[idx]$ in $vowels$ then
            $syllables \leftarrow syllables + 1$
        end if

        if $idx > 0$ then
            $previous\_char \leftarrow word[idx - 1]$
        else
            $previous\_char \leftarrow \emptyset$
        end if

        if $idx < word\_length - 1$ and $word[idx$ to $idx + 2]$ in $diphthongs$ then
            $syllables \leftarrow syllables - 1$
        end if

        if $idx < word\_length - 2$ and $word[idx$ to $idx + 3]$ in $triphthongs$ then
            $syllables \leftarrow syllables - 1$
        end if
    end for

    return $syllables$
end procedure
```

$$206.835 - 1.015 \times \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \times \left(\frac{\text{total syllables}}{\text{total words}} \right) \tag{4.6}$$

$$0.39 \times \left(\frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \times \left(\frac{\text{total syllables}}{\text{total words}} \right) - 15.59 \tag{4.7}$$

It is important to note that the Flesch Reading Ease formula was specifically designed for evaluating English texts. Nevertheless, Moreno et al. [75] conducted a study that adapted several readability formulas to accommodate the Portuguese language. Equation (4.8) and equation (4.9) show the formulas for the Flesch reading ease and Flesch-Kincaid grade level, respectively, modified to the Portuguese language.

$$227 - 1.04 \times \left(\frac{\text{total words}}{\text{total sentences}} \right) - 72 \times \left(\frac{\text{total syllables}}{\text{total words}} \right) \tag{4.8}$$

$$0.36 \times \left(\frac{\text{total words}}{\text{total sentences}}\right) + 10.4 \times \left(\frac{\text{total syllables}}{\text{total words}}\right) - 18 \tag{4.9}$$

4.5.1.2 Gunning Fog Index

Another widely used index is the Gunning Fog index, developed by Robert Gunning in 1952 [77]. To estimate its readability, this index relies on the average sentence length and the percentage of complex words in a text. Gunning defines complex words as words that have three or more syllables. Proper nouns, familiar jargon, or compound words are not considered complex words. The resulting score reflects the years of education required to understand the text. Equation (4.10) shows the formula for the Gunning Fog index.

$$0.4 \times \left(\frac{\text{total words}}{\text{total sentences}}\right) + 40 \times \left(\frac{\text{complex words}}{\text{total words}}\right) \tag{4.10}$$

Moreno et al. [75] modified the formula to the Portuguese context and got the values in equation (4.11).

$$0.49 \times \left(\frac{\text{total words}}{\text{total sentences}}\right) + 19 \times \left(\frac{\text{complex words}}{\text{total words}}\right) \tag{4.11}$$

4.5.1.3 Coleman-Liau Index

The Coleman-Liau Index, proposed by Meri Coleman and T. L. Liau in 1975 [78], measures readability based on the average number of characters per 100 words and the average number of sentences per 100 words. It provides a readability score corresponding to the U.S. grade level required for comprehension. Equation (4.12) shows the original Coleman-Liau index formula, and equation (4.13) shows the modified formula for Portuguese texts [75].

$$0.0588 \times \left(\frac{\text{total characters}}{\text{total sentences}} \times 100\right) - 29.6 \times \left(\frac{\text{total sentences}}{\text{total words}} \times 100\right) - 15.8 \tag{4.12}$$

$$0.054 \times \left(\frac{\text{total characters}}{\text{total sentences}} \times 100\right) - 21 \times \left(\frac{\text{total sentences}}{\text{total words}} \times 100\right) - 14 \tag{4.13}$$

4.5.1.4 Automated Readability Index (ARI)

The Automated Readability Index (ARI) is a widely used formula for assessing the readability of a text [79]. The ARI calculates the readability score based on the average number of characters per word and the average number of words per sentence. The resulting score indicates the grade level required to comprehend the text. A higher ARI score corresponds to a higher grade level, meaning more complex and challenging reading material.

$$4.71 \times \left(\frac{\text{total characters}}{\text{total words}}\right) + 0.5 \times \left(\frac{\text{total words}}{\text{total sentences}}\right) - 21.43 \tag{4.14}$$

$$4.6 \times \left(\frac{\text{total characters}}{\text{total words}}\right) + 0.44 \times \left(\frac{\text{total words}}{\text{total sentences}}\right) - 20 \tag{4.15}$$

4.6 Future Work

The previous analysis raised apprehensions regarding data size for objectivity and political stance detection. In addition, readability detection is lying on an algorithm that, despite being very accurate, can be further enhanced.

To address the first issue, this research proposes improving the data size and quality for objectivity and political stance methods. Starting with objectivity detection, although the dataset used in this study is expert-annotated, its limited size has resulted in only satisfactory performance. Gathering a more extensive and diverse dataset is crucial to achieve better results and enhance the model's objectivity detection. By collecting additional data, the model can be trained on more examples, leading to improved objectivity detection capabilities.

Moving on to political stance detection, the current dataset has certain limitations. It is based on parliamentary minutes and reflects the interventions of deputies in the Portuguese parliament. However, this writing style may differ significantly from a typical news article. Unfortunately, this dataset was the only available resource at the time. Therefore, future studies should construct a specific political stance dataset tailored to the Portuguese context. Building such a dataset will undoubtedly require substantial effort. Obtaining high-quality annotations can be accomplished through crowdsourcing or hiring experts via platforms like Amazon Turk, ensuring the annotations meet the necessary standards.

The second issue pertains to the algorithms used to estimate readability. As previously explained, in formulas where the number of syllables is required, there may be slight discrepancies between the estimated and actual number of syllables. This discrepancy arises because the syllables dataset does not contain all words, necessitating the estimation of syllable counts when a word is absent from the dataset. Currently, the algorithm achieves an accuracy of 82.66%. However, considering the extensive collection of samples within the syllables database, this research suggests that leveraging machine learning techniques to count syllables can further enhance the accuracy of syllable counting. By applying machine learning methods, it is expected that the accuracy of syllable counting, and subsequently the estimation of syllable counts, can be significantly improved. This improvement will contribute to more precise and reliable readability calculations.

4.7 Summary of the Chapter

This section explored approaches to analyzing news media articles in four dimensions: reliability, political stance, objectivity, and readability. Machine learning methods were employed and investigated for the first three dimensions, while established formulas were utilized for readability estimation.

For reliability classification, a BERT model with additional features was employed. The dataset, consisting of 124 thousand samples, was created by merging new data with a previous research dataset by Moura [80]. However, the dataset is unbalanced, with the majority class representing 88% of the samples. Additional features were derived from comprehensive data exploration, including sentiment analysis, entity count, and document topic. These features contributed to an impressive performance, achieving 0.9919 accuracy and 0.9811 macro F1 score, surpassing the results obtained in the previous research by Moura.

Due to the unavailability of news-related datasets for political stance classification, a political dataset based on interventions in the Portuguese parliament was used. This dataset was augmented using cutting-edge technology powered by OpenAI, resulting in an 8 thousand samples dataset. Despite extensive exploration, no significant insights were obtained. The model was trained using standard fine-tuning approaches and achieved an accuracy of 0.7777 and a macro F1 score of 0.7091. The previous results were obtained by training the model in five political categories (strict method). Additionally, the model was trained with a relaxed version, where political stance types were generalized to three, resulting in a 0.8022 accuracy and 0.7461 macro F1 score.

The objectivity detection task encountered a scarcity of data. To address this, a dataset annotated by experts and collected by Spinde et al. [38] was utilized for this dimension. The dataset was translated into Portuguese, and the model was trained using standard fine-tuning techniques. The training process resulted in an accuracy of 0.8260 and a macro F1 score of 0.7810, which can be considered satisfactory. However, to facilitate further development, it is essential to expand the available data. The current small dataset hinders the utilization of more powerful models that typically require more data to perform optimally. By increasing the dataset size, these models could be better utilized and potentially yield even better results.

A comprehensive exploration of various readability indices and formulas was conducted, some of which required syllable counting. However, counting syllables accurately poses a challenge, as currently, no algorithms are available for precise calculation. The resources provided by the Portal of the Portuguese Language were used to address this, where a dictionary containing words and their respective syllables was collected. This dictionary facilitated the syllable counting process in conjunction with the estimation algorithm. Additionally, Moreno et al. [75] made notable contributions by utilizing regression techniques to adapt established readability formulas specifically for the Portuguese language. This adaptation significantly helps estimate the readability of Portuguese texts, enhancing the overall assessment of their readability levels.

In conclusion, there is ample opportunity for improvement, particularly regarding the available data. As previously mentioned, the data is limited, highlighting the need for further collection efforts. However, despite this limitation, all the developed models have demonstrated either excellent or satisfactory performance. This reliability in their results instills confidence in using these models for future classification tasks on unseen data. It emphasizes the potential for further advancements and applications in the field with enhanced data availability and continuous improvement of the models.

Chapter 5

Prototype

5.1 Introduction 65
5.2 Use Cases 66
5.2.1 List of News Articles, the First Interaction 66
5.2.2 News Article Page, the Detailed Interaction 67
5.2.3 Chart, the Interactive Interaction 68
5.2.4 List Collections and Models, the Informative Interaction 71
5.3 Architecture 71
5.3.1 Backend 72
5.3.2 User Interface 75
5.4 Future Work 75
5.5 Summary of the Chapter 76

This chapter provides a broad overview of the development and design process of the prototype. The chapter begins by introducing the prototype and outlining its objective. Thereafter, the chapter delves into the available use cases within the prototype. Following this, the technical architectural design of the prototype is explained in detail, highlighting the underlying design and components. Lastly, the chapter ceases by debating future work and potential areas of improvement for the prototype, positioning the step for further development and enhancements.

5.1 Introduction

The previous chapter introduced an automatic approach to identifying axes that aid news consumers in understanding media news. However, this approach generates raw results from model predictions. The purpose of this chapter is to present the development of a user-friendly prototype that allows news consumers to access and interpret this information easily. This prototype is a website and thus is browser-based and does not require any prerequisites or installation.

Since this is a prototype, it is not yet ready for production use and lacks features. It serves as a proof-of-concept for this research. One major limitation is that it relies on a collection of pre-installed news articles rather than allowing online searches for articles. In other words, users cannot search for news articles in real-time using the prototype. Currently, the prototype offers a collection of COVID-19 news articles collected using the data collection tool (see Appendix A). These articles were then classified using the models described in the previous Chapter.

As mentioned earlier in Chapter 3, the development of this prototype was influenced by well-established and influential tools used for analyzing English written news, such as AllSides, The Factual, and Ad Fontes Media. The primary objective of the prototype is to provide news consumers with a tool to view bias-related information in news articles, enabling them to analyze bias along multiple dimensions. To accomplish this objective, the prototype will incorporate the most effective ideas from the aforementioned systems. The following list highlights the key features and insights gained from these well-established tools:

- **Straightforward Information**: The prototype will present information clearly and concisely, allowing users to understand the assigned value for each axis quickly. AllSides influence this approach;
- **Detailed Axis Information**: Users will have the ability to access detailed information about each axis, providing a more comprehensive and extensive classification result. The inspiration for this feature comes from The Factual.
- **Interactive Presentation**: For multi-class classification, a radar chart will present summarized information for each news outlet, allowing users to compare them. In the case of regression or binary classification, a 2D chart will be utilized to plot news articles and the centroid of each news outlet. This idea draws inspiration from Ad Fontes Media;

At the present date, the prototype is running on a Google Cloud on a `e2-standard-2` virtual machine, which consists of a 2 vCPU and 8GB of RAM virtual machine. The prototype is accessible via indice-media-portugues.pt and the API via indice-media-portugues.pt/v1/api.

5.2 Use Cases

This section describes the use cases of the prototype. It gives a brief look at the implementation of the key features and how a user can interact with them.

5.2.1 List of News Articles, the First Interaction

Upon accessing the website, users will receive a list of news articles. By default, these articles will be displayed in descending order based on their publication date. The list will be divided into pages, with 10 articles per page.

To enhance the browsing experience, users will have the option to apply filters to narrow down the displayed news articles. They can filter articles based on specific media outlets, select a specific date range, and even sort the articles in ascending or descending order based on their publication date.

This page plays a critical role as it is the initial interaction point between users and news articles. At the top of each news article's title, a list of values corresponding to each axis is displayed. These values represent either the class with the highest probability for the classification models or, in the case of regression models, a simple numeric value. This immediate exposition of values allows users to quickly learn the framing of each news article based on the determined axes.

Figure 5.1 provides a glimpse of the landing page prototype. The user is presented with a list of 10 news articles, and the first article is highlighted within a red square in the figure. Each news article card contains relevant information regarding its framing within the defined axes. Notably, the blue square highlights the labels assigned to each axis, offering additional context. Positioned on the left-hand side of the website is a filtering and sorting menu, identified with the green square, granting readers the ability to perform various operations on the displayed news articles. This feature empowers users to refine their browsing experience according to their preferences and requirements.

5.2.2 News Article Page, the Detailed Interaction

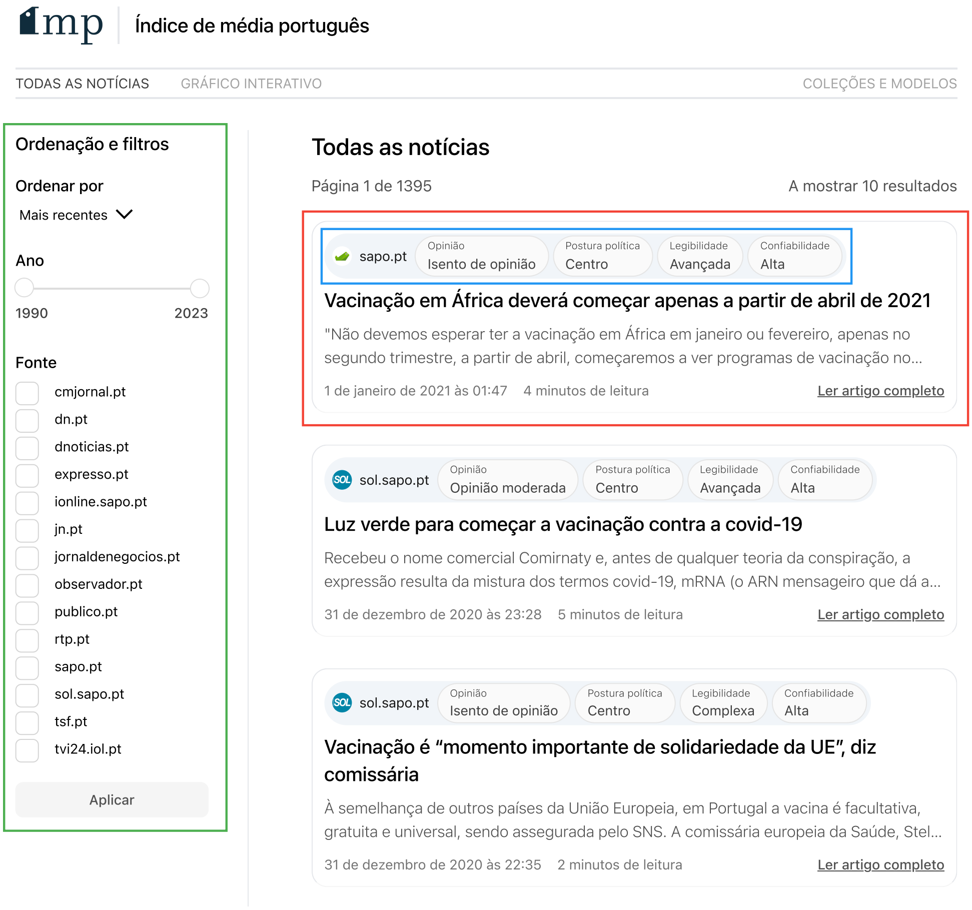
Upon selecting a news article from the list, the user will be referred to a dedicated page for that article. This page mimics the content of the original news piece, excluding any coexisting images. However, it provides a button that allows the user to visit the actual original news article if desired. Clicking on this button will redirect the user to the article in the news article's outlet website.

On this dedicated page, the user can view the text content of the news article. Additionally, on the right-hand side of the page, a description of the article's framing as resolved by the models is displayed. For each axis, a box will be shown, showcasing the assigned label. By default, the detailed information related to each axis will be hidden, with the intention of avoiding overwhelming the user with too much information. However, the user will have the option to expand each box to disclose detailed information about the article's framing.

In the case of multi-class classification, the probabilities associated with each class will be displayed. If the result is a regression value, the indicated value will fall within the range of the regression model values, if applicable.

Figure 5.2 shows the individual page of the news article "Vaccination in Africa should only start from April 2021" published on Sapo[1] on the 1st of January 2021. The content of the news article is displayed alongside its title, media outlet, published date, and the reading time of the content. On the top right corner, there is a button, highlighted in red, to read the original news article from the outlet's website. Highlighted in green, there is information about the framing. The label assigned to each axis is highlighted in bold to be easily readable. Upon expanding these

[1] sapo.pt



imp | Índice de média português

TODAS AS NOTÍCIAS GRÁFICO INTERATIVO COLEÇÕES E MODELOS

Ordenação e filtros

Ordenar por

Mais recentes

Ano

1990 2023

Fonte

□ cmjornal.pt
□ dn.pt
□ dnoticias.pt
□ expresso.pt
□ ionline.sapo.pt
□ jn.pt
□ jornaldenegocios.pt
□ observador.pt
□ publico.pt
□ rtp.pt
□ sapo.pt
□ sol.sapo.pt
□ tsf.pt
□ tvi24.iol.pt

Aplicar

Todas as notícias

Página 1 de 1395 A mostrar 10 resultados

sapo.pt | Opinião: Isento de opinião | Postura política: Centro | Legibilidade: Avançada | Confiabilidade: Alta

Vacinação em África deverá começar apenas a partir de abril de 2021

"Não devemos esperar ter a vacinação em África em janeiro ou fevereiro, apenas no segundo trimestre, a partir de abril, começaremos a ver programas de vacinação no...

1 de janeiro de 2021 às 01:47 4 minutos de leitura Ler artigo completo

sol.sapo.pt | Opinião: Opinião moderada | Postura política: Centro | Legibilidade: Avançada | Confiabilidade: Alta

Luz verde para começar a vacinação contra a covid-19

Recebeu o nome comercial Comirnaty e, antes de qualquer teoria da conspiração, a expressão resulta da mistura dos termos covid-19, mRNA (o ARN mensageiro que dá a...

31 de dezembro de 2020 às 23:28 5 minutos de leitura Ler artigo completo

sol.sapo.pt | Opinião: Isento de opinião | Postura política: Centro | Legibilidade: Complexa | Confiabilidade: Alta

Vacinação é "momento importante de solidariedade da UE", diz comissária

À semelhança de outros países da União Europeia, em Portugal a vacina é facultativa, gratuita e universal, sendo assegurada pelo SNS. A comissária europeia da Saúde, Stel...

31 de dezembro de 2020 às 22:35 2 minutos de leitura Ler artigo completo

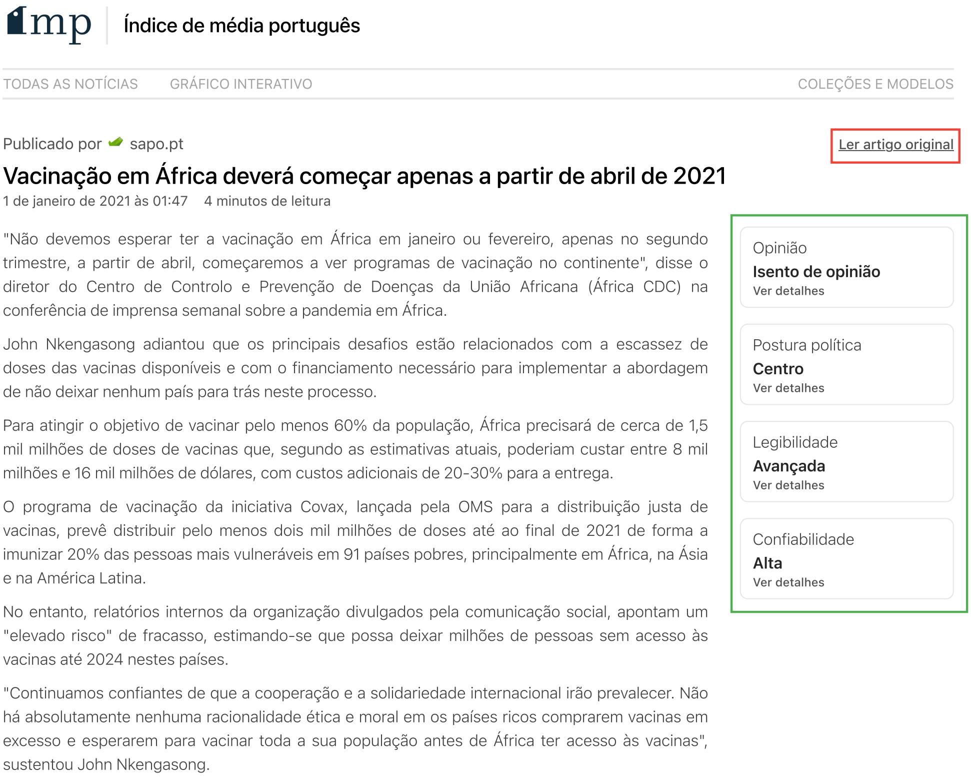
Figure 5.1: Main page of the web prototype.

options, the reader can see either the probabilities associated with each class or the position of the framing in a range, as seen in the green square in Figure 5.3

5.2.3 Chart, the Interactive Interaction

In addition to the features mentioned earlier, users can interact with the predicted data through charts. The prototype provides two types of charts: Radar and 2D. The Radar chart is suitable for visualizing multi-classification axes, making it particularly useful for dimensions like objectivity and political stance. On the other hand, the bi-dimensional chart is employed for binary classification axes, regression analysis, or formulas. This diverse range of chart options enables users to explore and analyze the predicted data from different perspectives, fostering a deeper understanding of the underlying patterns and relationships within the dataset.

It is also possible to obtain a linear representation for classification models. By utilizing Algorithm 2, a linear value within a specified range can be derived. This linear value serves as an approximation, allowing news readers to project classification values onto a continuous range.



Publicado por sapo.pt

Vacinação em África deverá começar apenas a partir de abril de 2021

1 de janeiro de 2021 às 01:47 4 minutos de leitura

"Não devemos esperar ter a vacinação em África em janeiro ou fevereiro, apenas no segundo trimestre, a partir de abril, começaremos a ver programas de vacinação no continente", disse o diretor do Centro de Controlo e Prevenção de Doenças da União Africana (África CDC) na conferência de imprensa semanal sobre a pandemia em África.

John Nkengasong adiantou que os principais desafios estão relacionados com a escassez de doses das vacinas disponíveis e com o financiamento necessário para implementar a abordagem de não deixar nenhum país para trás neste processo.

Para atingir o objetivo de vacinar pelo menos 60% da população, África precisará de cerca de 1,5 mil milhões de doses de vacinas que, segundo as estimativas atuais, poderiam custar entre 8 mil milhões e 16 mil milhões de dólares, com custos adicionais de 20-30% para a entrega.

O programa de vacinação da iniciativa Covax, lançada pela OMS para a distribuição justa de vacinas, prevê distribuir pelo menos dois mil milhões de doses até ao final de 2021 de forma a imunizar 20% das pessoas mais vulneráveis em 91 países pobres, principalmente em África, na Ásia e na América Latina.

No entanto, relatórios internos da organização divulgados pela comunicação social, apontam um "elevado risco" de fracasso, estimando-se que possa deixar milhões de pessoas sem acesso às vacinas até 2024 nestes países.

"Continuamos confiantes de que a cooperação e a solidariedade internacional irão prevalecer. Não há absolutamente nenhuma racionalidade ética e moral em os países ricos comprarem vacinas em excesso e esperarem para vacinar toda a sua população antes de África ter acesso às vacinas", sustentou John Nkengasong.

Opinião
Isento de opinião
Ver detalhes

Postura política
Centro
Ver detalhes

Legibilidade
Avançada
Ver detalhes

Confiabilidade
Alta
Ver detalhes

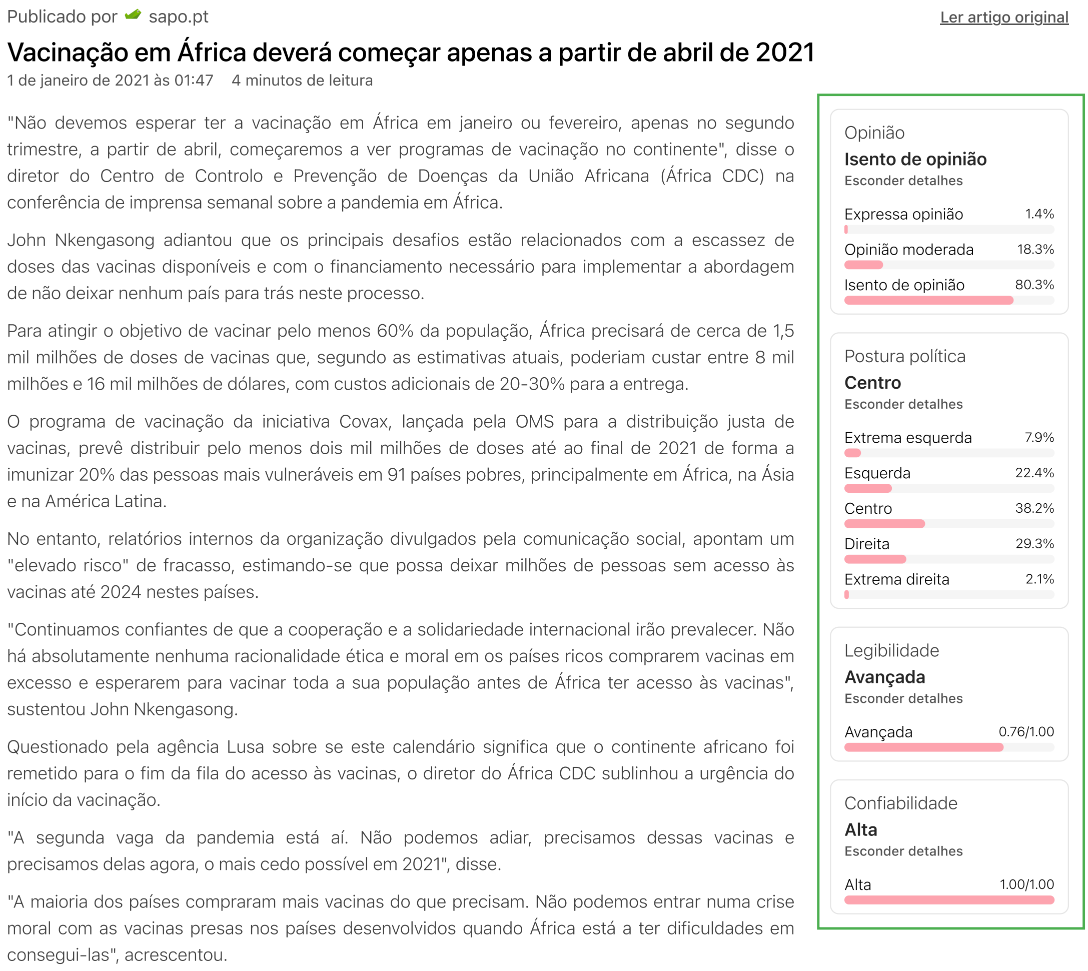
Figure 5.2: Individual page of a news article in the web prototype.

However, it is important to note that users are presented with a disclaimer explicitly stating that the classification model values are approximations of linear values.

In classification problems, the model produces a list of probabilities for each class rather than a continuous value within a range. The model is trained to classify samples into specific classes without differentiating between two samples within the same class. On the other hand, in regression problems, samples are distributed along a continuous range of values. Therefore, the algorithm estimates a value that can be observed in a 2D chart. It is crucial to emphasize that this algorithm provides an approximation, and readers are explicitly notified of this approximation through the provided disclaimer. This ensures readers know the limitations and potential discrepancies when interpreting the projected linear values from the classification model.

Figure 5.4 shows a radar chart example using the objectivity axis. On the left, in the green square, the user can select the model and filter with a year range and a list of collections. On the right, the user has the radar chart. The media outlets are selected with a search box with available media outlets.

Figure 5.5 provides a bi-dimensional chart illustrating the relationship between objectivity and reliability values. The X-axis represents the objectivity value, while the Y-axis represents the reliability value. On the left side of the chart, within the green square, users can find the model



Publicado por sapo.pt

Ler artigo original

Vacinação em África deverá começar apenas a partir de abril de 2021

1 de janeiro de 2021 às 01:47 4 minutos de leitura

"Não devemos esperar ter a vacinação em África em janeiro ou fevereiro, apenas no segundo trimestre, a partir de abril, começaremos a ver programas de vacinação no continente", disse o diretor do Centro de Controlo e Prevenção de Doenças da União Africana (África CDC) na conferência de imprensa semanal sobre a pandemia em África.

John Nkengasong adiantou que os principais desafios estão relacionados com a escassez de doses das vacinas disponíveis e com o financiamento necessário para implementar a abordagem de não deixar nenhum país para trás neste processo.

Para atingir o objetivo de vacinar pelo menos 60% da população, África precisará de cerca de 1,5 mil milhões de doses de vacinas que, segundo as estimativas atuais, poderiam custar entre 8 mil milhões e 16 mil milhões de dólares, com custos adicionais de 20-30% para a entrega.

O programa de vacinação da iniciativa Covax, lançada pela OMS para a distribuição justa de vacinas, prevê distribuir pelo menos dois mil milhões de doses até ao final de 2021 de forma a imunizar 20% das pessoas mais vulneráveis em 91 países pobres, principalmente em África, na Ásia e na América Latina.

No entanto, relatórios internos da organização divulgados pela comunicação social, apontam um "elevado risco" de fracasso, estimando-se que possa deixar milhões de pessoas sem acesso às vacinas até 2024 nestes países.

"Continuamos confiantes de que a cooperação e a solidariedade internacional irão prevalecer. Não há absolutamente nenhuma racionalidade ética e moral em os países ricos comprarem vacinas em excesso e esperarem para vacinar toda a sua população antes de África ter acesso às vacinas", sustentou John Nkengasong.

Questionado pela agência Lusa sobre se este calendário significa que o continente africano foi remetido para o fim da fila do acesso às vacinas, o diretor do África CDC sublinhou a urgência do início da vacinação.

"A segunda vaga da pandemia está aí. Não podemos adiar, precisamos dessas vacinas e precisamos delas agora, o mais cedo possível em 2021", disse.

"A maioria dos países compraram mais vacinas do que precisam. Não podemos entrar numa crise moral com as vacinas presas nos países desenvolvidos quando África está a ter dificuldades em consegui-las", acrescentou.

Opinião

Isento de opinião

Esconder detalhes

Expressa opinião 1.4%

Opinião moderada 18.3%

Isento de opinião 80.3%

Postura política

Centro

Esconder detalhes

Extrema esquerda 7.9%

Esquerda 22.4%

Centro 38.2%

Direita 29.3%

Extrema direita 2.1%

Legibilidade

Avançada

Esconder detalhes

Avançada 0.76/1.00

Confiabilidade

Alta

Esconder detalhes

Alta 1.00/1.00

Figure 5.3: Individual page of a news article in the web prototype. The framing detailed are expanded in the green square box.

selection, axis range, and filtering options. Meanwhile, the chart is on the right side, enclosed in the red square. The centroid of each media outlet, represented by their respective logos, is displayed within the chart.

By clicking on a specific media outlet logo within the orange square, users can visualize all the framed news articles associated with that outlet. Hovering over a data point representing a news article reveals information about its corresponding X and Y values, as shown in the blue box. Clicking on a data point redirects the user to a dedicated page for that particular article.

At the bottom of the chart, enclosed within the purple box, a disclaimer is presented. This disclaimer informs users that the values obtained from the classification models are approximations, as explained previously. It ensures transparency by acknowledging the limitations of the classification models utilized in the platform.

Algorithm 2 Algorithm to convert a list of probabilities into a continuous value between -1 and 1.

```
procedure PROBABILITIES_TO_VALUE($probabilities$)
    $number\_of\_classes \leftarrow length(probs)$
    $half \leftarrow number\_of\_classes/2$
    $step \leftarrow 1$
    $indices \leftarrow$ [values from $-half$ to $+half$ with steps of $step$]

    $weighted\_indices \leftarrow indices \cdot probabilities$
    $weighted\_value \leftarrow sum(weighted\_indices)$

    $normalized\_value \leftarrow weighted\_value/half$
    return $normalized\_value$
end procedure
```

5.2.4 List Collections and Models, the Informative Interaction

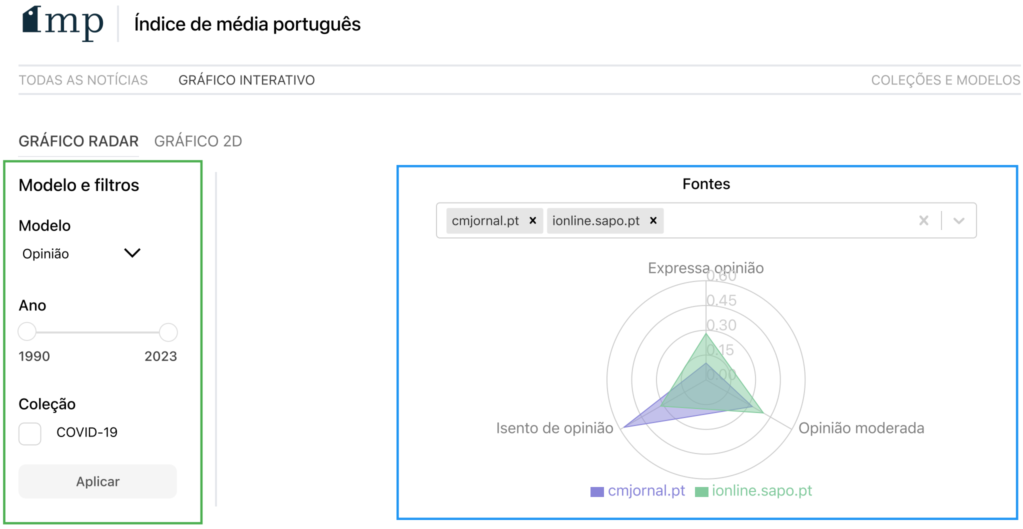
Users can conveniently check the available collections and models within the system. They are provided with a clear description of each collection or model, outlining its purpose and composition. In the case of models, users can easily verify the specific task that a model performs, explore the classes and their respective ranges (where applicable), and even download the model directly from the prototype with a button.

This page is a comprehensive resource for obtaining detailed information about the news articles accessible on the website, including how they were framed. It empowers news consumers by informing them about the potential limitations associated with each collection and model. Moreover, users are warned about possible issues or considerations they should consider while examining the results, ensuring they approach the data critically.

Figure 5.6 shows the collections and models page. Highlighted in the green square box, the information about a collection is available. In the example, the collection is a COVID-19 collection of articles collected from Portuguese media outlets and on the right, highlighted in the blue square. It displays its classes or ranges and a brief description.

5.3 Architecture

The architecture of this prototype can be divided into two main components: the prototype UI and the backend. A visual representation of the architecture, including its components and their interactions, is depicted in Figure 5.7. This diagram provides a clear overview of how the prototype is structured and how information flows between the UI and the backend. It serves as a helpful reference for understanding the underlying design and functionality of the system.



Modelo: Opinião

Ano: 1990 – 2023

□ COVID-19

Fontes: cmjornal.pt ×, ionline.sapo.pt ×

Figure 5.4: Radar chart example with two media outlets in the web prototype.

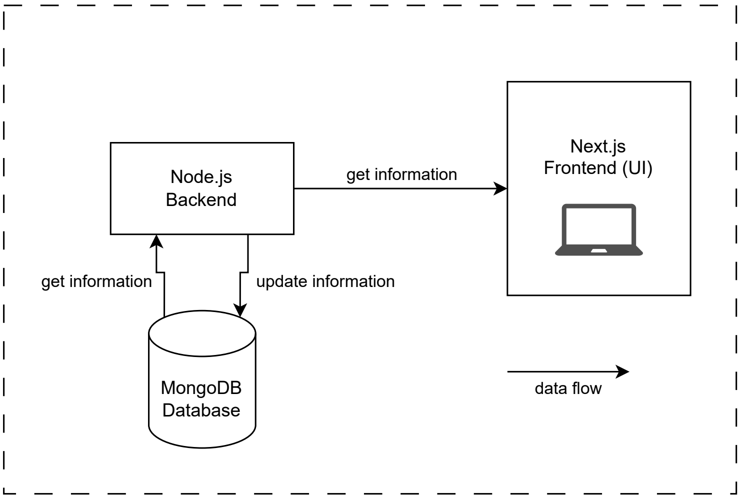
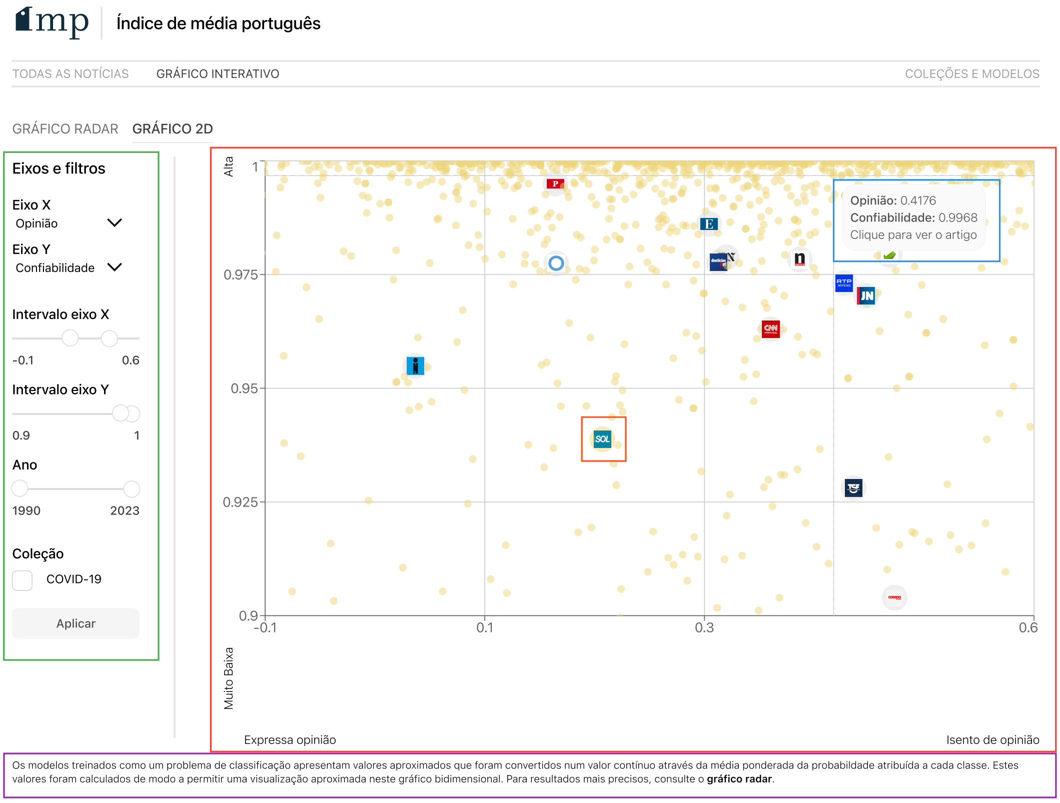


Figure 5.7: Architecture of the prototype.

The backend and the frontend are isolated in a Docker container, and the website is exposed in the server with NGINX, acting as a reverse proxy. Inside the containerized system, the backend runs on port 3001, and the frontend in port 3000. Then, the NGINX server is responsible for making the UI available in ports 80 and 433 (HTTP and HTTPS, respectively) and making the API available in the path `/v1/api`.

5.3.1 Backend

The backend of the prototype runs on a Docker container. It is a REST API written in Node.js and Express.js. This component is responsible for interacting with the database and providing



Eixo X
Opinião

Eixo Y
Confiabilidade

□ COVID-19

Opinião: 0.4176
Confiabilidade: 0.9968

Os modelos treinados como um problema de classificação apresentam valores aproximados que foram convertidos num valor contínuo através da média ponderada da probabildade atribuída a cada classe. Estes valores foram calculados de modo a permitir uma visualização aproximada neste gráfico bidimensional. Para resultados mais precisos, consulte o **gráfico radar**.

Figure 5.5: Bi-dimensional chart example in the web prototype.

information via an interface to the exterior. It uses a REST architecture on the HTTP protocol. This enables the UI to make calls to request information.

It supports multiple endpoints to create, read, delete, and update objects in the database. The information stored in the database consists of the articles, the collections, the models, and the article predictions. The article object contains information about the news article's content, title, publish date, and original link. It also holds a reference to its collection id. An article cannot exist without a collection. The collection contains information about its name and description. A model has information about its configuration, name, description, and a reference to its binary file. Finally, the article predictions objects hold a value and a connection to its article and model.

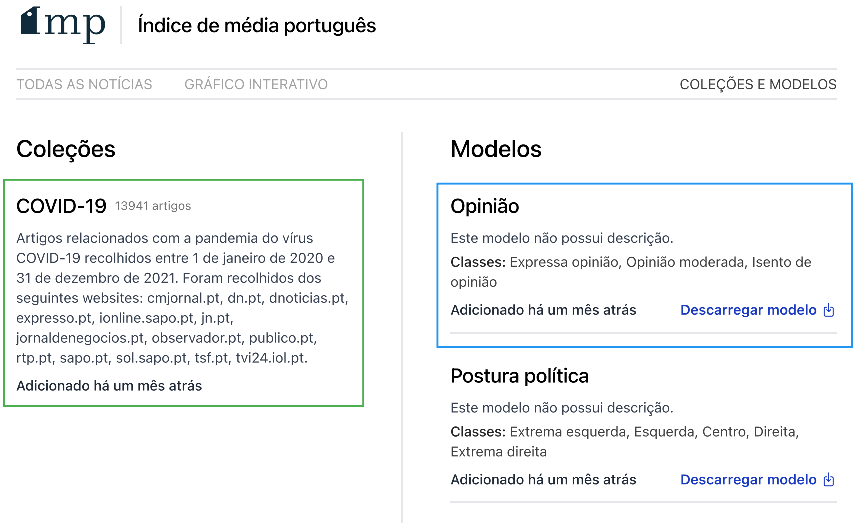
Table 5.1 shows the available endpoints and the method used. The complete endpoint specification is in a Swagger format in the prototype source code[2].

The UI does not implement authentication for these endpoints, as this prototype is a proof of concept. In addition, some endpoints do not fit a UI implementation and are only accessible via manual request (*e.g.*, `curl`, Postman, etc.). Table 5.1 also shows which endpoints have a corresponding UI or are requested in the prototype UI.

[2] github.com/luist18/dissertation-platform

Table 5.1: Backend's REST API endpoints, their description, and whether they have an implemented UI in the prototype.

Endpoint	**Description**	**UI**
`GET` /	Check the server status	No
`GET` /article	Get all articles	**Yes**
`GET` /article/<id>	Get article by id	**Yes**
`GET` /collection	Get all collections	**Yes**
`POST` /collection	Create new collection	No
`GET` /collection/<id>	Get collection by id	No
`PUT` /collection/<id>	Update collection by id	No
`DELETE` /collection/<id>	Delete collection by id	No
`POST` /collection/<id>/articles	Create news articles in a collection by collection id	No
`DELETE` /collection/<id>/articles	Delete news articles in a collection by collection id	No
`GET` /model	Get all models	**Yes**
`POST` /model	Create new model	No
`GET` /model/<id>	Get model by id	No
`PUT` /model/<id>	Update model by id	No
`DELETE` /model/<id>	Delete model by id	No
`GET` /predictions	Get all predictions	**Yes**



Coleções

COVID-19 13941 artigos

Artigos relacionados com a pandemia do vírus COVID-19 recolhidos entre 1 de janeiro de 2020 e 31 de dezembro de 2021. Foram recolhidos dos seguintes websites: cmjornal.pt, dn.pt, dnoticias.pt, expresso.pt, ionline.sapo.pt, jn.pt, jornaldenegocios.pt, observador.pt, publico.pt, rtp.pt, sapo.pt, sol.sapo.pt, tsf.pt, tvi24.iol.pt.

Adicionado há um mês atrás

Modelos

Opinião

Este modelo não possui descrição.

Classes: Expressa opinião, Opinião moderada, Isento de opinião

Adicionado há um mês atrás

Postura política

Este modelo não possui descrição.

Classes: Extrema esquerda, Esquerda, Centro, Direita, Extrema direita

Adicionado há um mês atrás

Figure 5.6: Collections and models page in the web prototype.

5.3.2 User Interface

The user interface is implemented with Next.js and uses both client-side and server-side rendering. It supports multiple pages, as previously described in the use cases. The interface is intended to be as simple and plain as possible to allow a swift interaction between itself and the news consumers. It offers general features for data listing, such as sorting, filtering, and pagination.

The charts in the interface are rendered using the Rechart.js library. It was the library that better suited the needs. However, it has some limitations on performance and customization. Future improvements should consider building a chart library only for this prototype.

5.4 Future Work

This Chapter presents the initial version of a prototype for a Portuguese media bias platform. As indicated by its prototype status, there is significant room for improvement. This research proposes three key features to focus on for further development: grouping similar or related news articles (identical to the approach taken by AllSides), real-time search capability, and the integration of community feedback.

Grouping similar news articles side by side would empower news readers to compare coverage across multiple sources, enabling them to identify potential biases or select the article that aligns best with their preferences.

Real-time search functionality is crucial as it would allow news readers to frame any news article according to their specific search criteria. Currently, the prototype only supports pre-defined collections. This limitation stems from the absence of a suitable server that can process news articles in real time. By implementing real-time search, users would have a more dynamic and customizable experience.

Lastly, including community feedback is vital for enhancing the platform's models and validating their performance. In a platform aimed at informing news consumers, gathering feedback from the community plays a pivotal role. By continuously collecting feedback, efforts can be directed toward addressing issues that may affect users' experience or result in misleading outcomes. This feedback loop ensures ongoing improvement and promotes transparency and trust within the platform.

5.5 Summary of the Chapter

This chapter presents the media bias platform prototype, built upon the four axes developed in this study. The prototype is a user-friendly web platform inspired by prominent platforms used by U.S. news outlets.

The primary purpose of the prototype is to provide readers with a seamless experience, allowing them to explore how news articles are framed easily. Users can list news articles, apply filters, and sort them according to their preferences. Additionally, they can access detailed versions of individual news articles through dedicated pages. The website also incorporates interactive charts, enabling users to visually compare news articles and outlets with other articles or outlets more conveniently.

The system's architecture is designed to be straightforward. The frontend of the platform sends requests to the backend, which is responsible for managing the information in the database and handling those requests. The relevant information is made available through a REST API.

Since this prototype is presented as a work in progress, there is ample room for improvement. The ultimate goal is to transform this prototype into a fully functional solution tailored to the specific needs of the Portuguese news environment.

Chapter 6

Conclusions

6.1 Summary . . . 77
6.2 Threats to Validity . . . 78
6.3 Contribution . . . 79
6.4 Future Work . . . 80

This chapter provides the dissertation summary outlining the research's threats to validity, crucial points, and future work.

6.1 Summary

This research has done a study on media bias in the Portuguese domain. The nature and roots of media bias were investigated, revealing many forms of media bias. This research focused on media bias by word choice and labeling form. The methods to detect this bias have been involved in the last decades. Initially, the detection was done manually using codebooks. This method still exists at the present date. More recently, computer scientists have proposed algorithms and artificial intelligence models to analyze text and classify sentences. This approach was found useful in the aspect of media bias classification and detection.

Considering that it is possible to investigate and classify textual features in texts. The next step focused on identifying axes of interest to help news readers understand news articles and their nature. The research identified four dimensions of interest: reliability, political stance, objectivity, and readability. Determining the reliability and credibility of a media outlet is crucial. It plays a vital role in establishing the trust readers can place in the publication's information. The axis of political stance assesses where news articles stand on the political spectrum. It aims to identify and describe the political orientation or bias present in the articles, allowing readers to understand their perspectives and leanings. On the other hand, objectivity examines the absence of personal opinions in news articles. It emphasizes presenting information factually, free from subjective views, biases, or personal beliefs. Lastly, the readability axis focuses on how articles are written and evaluate their level of comprehension for readers. By defining and addressing these four axes, this chapter establishes a comprehensive framework that enables the analysis and interpretation of news articles from multiple dimensions, encompassing reliability, political stance, objectivity, and readability.

Next, the research delves into describing the models used for each axis. The methodology employed consisted of fine-tuning BERT models or applying established metrics. For axes such as reliability, political stance, and objectivity, machine learning models were employed. The standard approach involved data analysis, model development, and model evaluation. It is important to note that some models featured more complex architecture, primarily due to the inclusion of engineered features derived from the data. However, a notable challenge faced by these machine learning models was the scarcity of data, which limited the ability to gather sufficient insights for model improvement. Regarding the readability axis, literature metrics were employed to assess the readability of news articles. These metrics were adapted and customized to cater to the specific requirements of the Portuguese language.

Once the groundwork with the models was completed, a prototype was developed to present the predicted data from these models in a user-friendly manner. The development process involved analyzing existing prominent platforms to gather valuable and established concepts and ideas. Based on this research, a prototype was created, incorporating general features aimed at visualizing the framing of news articles by the models. The prototype offers various functionalities that enable news readers to comprehend the framing of articles and compare media outlets and news articles based on the defined axes. These features empower users to gain insights into how news articles are positioned and categorized, facilitating a comprehensive understanding of the information presented.

In conclusion, the research successfully achieved its objectives by identifying specific dimensions of media bias and proposing innovative solutions within the context of the Portuguese language. This significant work lays the foundation for future development and research in this field. Furthermore, the prototype developed during the research represents a significant milestone and offers tremendous potential for future advancements. It provides a platform for further exploration, allowing researchers to refine and enhance its features and contribute to addressing the issue of media bias. The prototype catalyzes ongoing efforts to tackle this complex problem and offers valuable opportunities for improvement and expansion in the future.

6.2 Threats to Validity

This section discusses potential threats to the validity of the results and conclusions drawn in this dissertation. Specifically, the threats arise from the dataset sizes used for training the models and the absence of benchmark results for model comparison.

Threats from Dataset Sizes The size of the dataset used for training the models can have implications for the generalizability and reliability of the findings. In this study, the following threats related to dataset sizes are identified:

1. Small Dataset Size: One potential threat is using relatively small datasets for training the models. Limited dataset sizes may not adequately capture the underlying patterns and complexities of the problem domain, leading to models that are not robust enough to handle real-world scenarios. The generalizability of the findings might be compromised due to overfitting or insufficient representation of the target population.
2. Imbalanced Dataset: Another threat arises from imbalanced datasets, where certain classes or categories are underrepresented compared to others. This can bias the models towards the majority class and result in suboptimal performance for minority classes. The lack of sufficient training examples for these classes may hinder the model's ability to predict their occurrences in real-world scenarios accurately.

Threats from Absence of Benchmark Results The absence of benchmark results against which to compare the performance of the developed models introduces the following threats:

1. Lack of Performance Benchmarks: Assessing the effectiveness and superiority of developed models becomes difficult without established benchmarks. A performance baseline is required to assess the practical significance of the proposed methods and measure the improvements they achieve. The lack of such a baseline makes it difficult to make firm claims about the contributions and advancements of the proposed models.
2. Difficulty in Comparative Analysis: Without benchmark results, conducting a comparative analysis between the developed models and existing approaches becomes difficult. Comparative studies are essential for determining various models' relative strengths and weaknesses and making informed decisions about their suitability for specific applications. The lack of benchmark results limits the ability to perform such analyses, potentially limiting understanding and progress in the field.

6.3 Contribution

All in all, it is possible to highlight the following contributions from this research:

Datasets Given the absence of suitable datasets for this research, a significant research focus was dedicated to sourcing, augmenting, translating, and creating datasets. The following datasets, presented in this work have been identified as valuable resources for future research. These datasets were either newly created or modified from existing ones to cater specifically to the research requirements:

- Reliability Arquivo.pt Dataset: consists of 20 thousand samples of reliable and unreliable Portuguese news articles. The news articles were retriMachine learning models were employed ford using multiple query terms from a specific date range.
- Large Reliability Dataset: it is the biggest reliability dataset for Portuguese news. It represents an augmented version of the dataset collection by Moura [80] and the dataset created in this research.
- BABE (Portuguese) Dataset: consists of the 3 700 samples of the BABE dataset translated to European Portuguese using the DeepL API. It is a dataset for bias and objectivity classification.
- Political stance dataset: based on a collection of interventions from Portuguese deputies in the Portuguese parliament, this dataset used the powerful GPT-3.5 model from OpenAI to augment its size and result in an 8 thousand samples dataset of political orientation labeled texts.

Data Collection Tool The data collection tool developed in this work contributes to future works as it provides an excellent means to collect and curate news articles. It allows retrieving news articles' content and metadata with a query from a specific set of websites within a range.

Prototype The prototype developed and described in this work presents an opportunity to further development and improvement. The prototype is a solid foundation for a Portuguese news framing website like the ones seen for U.S. news. The developed work is open-source and allows community development.

6.4 Future Work

As this research poses a novel introduction to the media bias problem in the Portuguese context, there is still an avenue to pursue with a few gaps to fill. Future work not only involves enhancing the axes models and detection but also should focus on identifying more axes of interest. As it presents a prototype for a media bias platform, it is also a starting point for future research and development.

Future research should focus on refining and enhancing existing models for the specified axes. This can involve exploring advanced machine learning algorithms, incorporating additional features, or fine-tuning the existing models based on feedback and evaluation results. By continually improving the models, their accuracy, robustness, and effectiveness can be enhanced, leading to more reliable and valuable outcomes.

The development of the prototype can be expanded in several areas to enhance its functionality. These include implementing real-time search, incorporating community feedback, and optimizing performance. The prototype's news articles are pre-installed or manually added through API requests. By implementing real-time search, users can stay updated on ongoing news events and easily access relevant information. This feature will provide an overview of search results for specific events of interest, enabling users to follow the progression of news stories in real time. To further improve the prototype, it is crucial to incorporate community feedback. This will allow users to report any issues they encounter while using the platform, which can be addressed promptly.

Additionally, gathering feedback from the community can help identify valuable features or enhancements that could be implemented to enhance the overall user experience. Lastly, it is essential to address the performance limitations of the prototype. Currently, the platform relies on heavy libraries, which can affect its performance. It is recommended that future development considers these cases and carefully considers whether it is more beneficial to create a streamlined solution that addresses the performance issues. This will ensure the prototype operates smoothly and efficiently, providing users with a seamless experience. By focusing on these aspects - implementing real-time search, incorporating community feedback, and optimizing performance - the prototype can be further developed to create a robust and user-friendly platform.

References

[1] Felix Hamborg. *Towards Automated Frame Analysis : Natural Language Processing Techniques to Reveal Media Bias in News Articles*. PhD thesis, Universität Konstanz, 2022. Accepted: 2022-03-22T06:15:47Z ISBN: 9781796226294. Cited on pages vii, viii, 1, 3, 7, 8, 9, 10, 11, and 12.

[2] Ad Fontes Media. Interactive Chart. Cited on pages vii, 14, 18, and 31.

[3] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 443–452, New York, NY, USA, 2009. Association for Computing Machinery. Cited on pages viii, 8, 9, and 14.

[4] Felix Hamborg, Norman Meuschke, Akiko Aizawa, and Bela Gipp. *Identification and Analysis of Media Bias in News Articles*. Humboldt-Universität zu Berlin, March 2017. Accepted: 2017-06-15T13:06:30Z. Cited on pages 1 and 18.

[5] Benjamin Parker Bass, Ian Fette, Paul Mans, Monica Seth, Jim Sullivan, and Paul Washburn. News Bias Explored. Cited on page 2.

[6] Matthew Gentzkow and Jesse M. Shapiro. Media Bias and Reputation. *Journal of Political Economy*, 114(2):280–316, April 2006. Publisher: The University of Chicago Press. Cited on pages 2 and 19.

[7] Stefano DellaVigna and Ethan Kaplan. The Fox News Effect: Media Bias and Voting*. *The Quarterly Journal of Economics*, 122(3):1187–1234, August 2007. Cited on pages 2, 6, 10, and 12.

[8] Tim Groseclose and Jeffrey Milyo. A Measure of Media Bias*. *The Quarterly Journal of Economics*, 120(4):1191–1237, November 2005. Cited on pages 2 and 10.

[9] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, March 2018. Publisher: American Association for the Advancement of Science. Cited on pages 2, 8, and 19.

[10] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, March 2018. Publisher: American Association for the Advancement of Science. Cited on page 2.

[11] Renée Hobbs. The Seven Great Debates in the Media Literacy Movement. *Journal of Communication*, 48(1):16–32, March 1998. Cited on page 2.

[12] Douglas Kellner and Jeff Share. Toward Critical Media Literacy: Core concepts, debates, organizations, and policy. *Discourse: Studies in the Cultural Politics of Education*, 26(3):369–386, September 2005. Publisher: Routledge _eprint: https://doi.org/10.1080/01596300500200169. Cited on page 2.

[13] Jakob-Moritz Eberl, Hajo G. Boomgaarden, and Markus Wagner. One Bias Fits All? Three Types of Media Bias and Their Effects on Party Preferences. *Communication Research*, 44(8):1125–1148, December 2017. Publisher: SAGE Publications Inc. Cited on pages 3, 6, and 8.

[14] Felix Hamborg, Karsten Donnay, and Bela Gipp. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415, December 2019. Cited on pages 3 and 18.

[15] Dave D'Alessio and Mike Allen. Media Bias in Presidential Elections: A Meta-Analysis. *Journal of Communication*, 50(4):133–156, December 2000. Cited on pages 6, 7, and 10.

[16] Tim Groeling. Media Bias by the Numbers: Challenges and Opportunities in the Empirical Study of Partisan News. *Annual Review of Political Science*, 16(1):129–151, 2013. _eprint: https://doi.org/10.1146/annurev-polisci-040811-115123. Cited on pages 6 and 8.

[17] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. A Computational Framework for Media Bias Mitigation. *ACM Transactions on Interactive Intelligent Systems*, 2(2):1–32, June 2012. Cited on page 7.

[18] Alden Williams. Unbiased Study of Television News Bias. *Journal of Communication*, 25(4):190–199, 1975. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-2466.1975.tb00656.x. Cited on pages 7 and 8.

[19] Ruibo Liu, Chenyan Jia, and Soroush Vosoughi. A Transformer-based Framework for Neutralizing and Reversing the Political Polarity of News Articles. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):65:1–65:26, 2021. Cited on page 8.

[20] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. Automatically Neutralizing Subjective Bias in Text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489, April 2020. Number: 01. Cited on pages 8 and 55.

[21] Kartikey Pant, Tanvi Dadu, and Radhika Mamidi. Towards Detection of Subjective Bias using Contextualized Word Embeddings. In *Companion Proceedings of the Web Conference 2020*, WWW '20, pages 75–76, New York, NY, USA, 2020. Association for Computing Machinery. Cited on pages 8 and 55.

[22] Benjamin D. Horne, Dorit Nevo, John O'Donovan, Jin-Hee Cho, and Sibel Adalı. Rating Reliability and Bias in News Articles: Does AI Assistance Help Everyone? *Proceedings of the International AAAI Conference on Web and Social Media*, 13:247–256, July 2019. Cited on pages 8 and 19.

[23] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017. Cited on page 8.

[24] Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1478–1484, Marseille, France, May 2020. European Language Resources Association. Cited on page 8.

[25] SHANTO IYENGAR and ADAM SIMON. News Coverage of the Gulf Crisis and Public Opinion: A Study of Agenda-Setting, Priming, and Framing. *Communication Research*, 20(3):365–383, June 1993. Publisher: SAGE Publications Inc. Cited on page 8.

[26] Stephen Earl Bennett. The Persian Gulf War's Impact on Americans' Political Information. *Political Behavior*, 16(2):179–201, 1994. Publisher: Springer. Cited on page 10.

[27] Felix Hamborg, Anastasia Zhukova, and Bela Gipp. Illegal Aliens or Undocumented Immigrants? Towards the Automated Identification of Bias by Word Choice and Labeling. In Natalie Greene Taylor, Caitlin Christian-Lamb, Michelle H. Martin, and Bonnie Nardi, editors, *Information in Contemporary Society*, volume 11420, pages 179–187. Springer International Publishing, Cham, 2019. Series Title: Lecture Notes in Computer Science. Cited on page 10.

[28] Camiel J. Beukeboom and Christian Burgers. Linguistic Bias. In *Oxford Research Encyclopedia of Communication*. Oxford University Press, July 2017. Cited on pages 10 and 11.

[29] Camiel J. Beukeboom. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. In *Social cognition and communication*, Sydney symposium of social psychology, pages 313–330. Psychology Press, New York, NY, US, 2014. Cited on page 11.

[30] Sammy Mngqosini. Morocco's World Cup win against Belgium triggers riots in Brussels, November 2022. Cited on page 11.

[31] Zizi Papacharissi and Maria de Fatima Oliveira. Affective News and Networked Publics: The Rhythms of News Storytelling on #Egypt. *Journal of Communication*, 62(2):266–282, April 2012. Cited on page 12.

[32] Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1195–1198, New York, NY, USA, 2010. Association for Computing Machinery. Cited on page 12.

[33] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia, July 2018. Association for Computational Linguistics. Cited on page 13.

[34] Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3):102505, May 2021. Cited on page 13.

[35] Christoph Hube and Besnik Fetahu. Neural Based Statement Classification for Biased Language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, January 2019. Cited on page 13.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, December 2017. arXiv:1706.03762 [cs]. Cited on pages 13 and 27.

[37] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv:1810.04805 [cs]. Cited on pages 13 and 27.

[38] Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. Neural Media Bias Detection Using Distant Supervision With BABE – Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, 2021. arXiv:2209.14557 [cs]. Cited on pages 13, 22, 55, and 63.

[39] Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. In Plain Sight: Media Bias Through the Lens of Factual Reporting, September 2019. arXiv:1909.02670 [cs]. Cited on page 13.

[40] Timo Spinde. An Interdisciplinary Approach for the Automated Detection and Visualization of Media Bias in News Articles. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 1096–1103, December 2021. ISSN: 2375-9259. Cited on page 13.

[41] Jan-David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela Gipp. A domain-adaptive pre-training approach for language bias detection in news. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–7, Cologne Germany, June 2022. ACM. Cited on page 13.

[42] T. Spinde, L. Rudnitckaia, K. Sinha, F. Hamborg, B. Gipp, and K. Donnay. MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics, May 2021. arXiv:2105.11910 [cs]. Cited on pages 13, 18, and 22.

[43] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to Fine-Tune BERT for Text Classification?, February 2020. arXiv:1905.05583 [cs]. Cited on page 13.

[44] Rukhma Qasim, Waqas Haider Bangyal, Mohammed A. Alqarni, and Abdulwahab Ali Almazroi. A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification. *Journal of Healthcare Engineering*, 2022:1–17, January 2022. Cited on page 13.

[45] Souneil Park, Minsam Ko, Jungwoo Kim, Ho-Jin Choi, and Junehwa Song. NewsCube 2.0: an exploratory design of a social news website for media bias mitigation. In *Workshop on Social Recommender Systems*, 2011. Cited on page 14.

[46] Polígrafo SIC. Polígrafo SIC. Cited on page 14.

[47] Observador. Fact Check – notícias, opinião, rádio, fotos e podcasts, February 2023. Cited on page 14.

[48] AllSides. AllSides Media Bias Chart, February 2019. Cited on pages 14 and 18.

[49] Felix Hamborg, Anastasia Zhukova, and Bela Gipp. Illegal Aliens or Undocumented Immigrants? Towards the Automated Identification of Bias by Word Choice and Labeling. In Natalie Greene Taylor, Caitlin Christian-Lamb, Michelle H. Martin, and Bonnie Nardi, editors, *Information in Contemporary Society*, Lecture Notes in Computer Science, pages 179–187, Cham, 2019. Springer International Publishing. Cited on page 18.

[50] The Factual - Unbiased News, Trending Topics - The Factual. Cited on pages 18 and 19.

[51] Ground News. Cited on page 18.

[52] Sebastian Kula, Michał Choraś, Rafał Kozik, Paweł Ksieniewicz, and Michał Woźniak. Sentiment Analysis for Fake News Detection by Means of Neural Networks. In Valeria V. Krzhizhanovskaya, Gábor Závodszky, Michael H. Lees, Jack J. Dongarra, Peter M. A. Sloot, Sérgio Brissos, and João Teixeira, editors, *Computational Science – ICCS 2020*, Lecture Notes in Computer Science, pages 653–666, Cham, 2020. Springer International Publishing. Cited on page 19.

[53] Miguel A. Alonso, David Vilares, Carlos Gómez-Rodríguez, and Jesús Vilares. Sentiment Analysis for Fake News Detection. *Electronics*, 10(11):1348, January 2021. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute. Cited on page 19.

[54] Michelle YoungJin Kim and Kristen Marie Johnson. CLoSE: Contrastive Learning of Subframe Embeddings for Political Bias Classification of News Media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2780–2793, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. Cited on pages 20 and 47.

[55] Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. We Can Detect Your Bias: Predicting the Political Ideology of News Articles, October 2020. arXiv:2010.05338 [cs]. Cited on pages 20 and 47.

[56] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning Subjective Language. *Computational Linguistics*, 30(3):277–308, September 2004. Cited on page 20.

[57] Andrés Montoyo, Patricio Martínez-Barco, and Alexandra Balahur. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, 53(4):675–679, November 2012. Cited on page 20.

[58] Gaël Dias, Dinko Lambov, and Veska Noncheva. High-level Features for Learning Subjective Language across Domains. *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1):199–202, March 2009. Number: 1. Cited on page 21.

[59] C. S. Richard Chan, Charuta Pethe, and Steven Skiena. Natural language processing versus rule-based text analysis: Comparing BERT score and readability indices to predict crowdfunding outcomes. *Journal of Business Venturing Insights*, 16:e00276, November 2021. Cited on page 21.

[60] Miguel Oliveira and Tiago Melo. Investigando Features de Sentenças para Classificação de Subjetividade e Polaridade em Português do Brasil. In *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 270–281. SBC, October 2020. ISSN: 2763-9061. Cited on page 22.

[61] Ricardo Moura, Rui Sousa-Silva, and Henrique Lopes Cardoso. Automated Fake News Detection Using Computational Forensic Linguistics. In Goreti Marreiros, Francisco S. Melo, Nuno Lau, Henrique Lopes Cardoso, and Luís Paulo Reis, editors, *Progress in Artificial Intelligence*, Lecture Notes in Computer Science, pages 788–800, Cham, 2021. Springer International Publishing. Cited on pages 22, 23, 38, and 39.

[62] Arquivo.pt - pesquise páginas do passado! Cited on page 24.

[63] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In Ricardo Cerri and Ronaldo C. Prati, editors, *Intelligent Systems*, Lecture Notes in Computer Science, pages 403–417, Cham, 2020. Springer International Publishing. Cited on pages 28 and 29.

[64] João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. Advancing Neural Encoding of Portuguese with Transformer Albertina PT-*, June 2023. arXiv:2305.06721 [cs]. Cited on pages 28, 29, and 57.

[65] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online, 2020. Association for Computational Linguistics. Cited on pages 28 and 29.

[66] João Pedro Baptista and Anabela Gradim. Online disinformation on Facebook: the spread of fake news during the Portuguese 2019 election. *Journal of Contemporary European Studies*, November 2020. Publisher: Routledge. Cited on pages 37 and 38.

[67] Miguel António Palma dos Santos Sozinho Ramalho. High-level Approaches to Detect Malicious Political Activity on Twitter. Master's thesis, Universidade do Porto, July 2020. Accepted: 2022-09-13T21:32:29Z. Cited on page 38.

[68] Paulo Pena. Fake news: sites portugueses com mais de dois milhões de seguidores, November 2018. Cited on page 38.

[69] Ricardo Moura, Rui Sousa-Silva, and Henrique Lopes Cardoso. Automated Fake News Detection Using Computational Forensic Linguistics. In Goreti Marreiros, Francisco S. Melo, Nuno Lau, Henrique Lopes Cardoso, and Luís Paulo Reis, editors, *Progress in Artificial Intelligence*, volume 12981, pages 788–800. Springer International Publishing, Cham, 2021. Series Title: Lecture Notes in Computer Science. Cited on page 38.

[70] Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks, June 2021. Cited on page 40.

[71] Francisco Varandas Soares Graça. A política e os media: o enviesamento da imprensa portuguesa em 2009 e 2015. Master's thesis, ISCTE-Instituto Universitário de Lisboa, 2017. Accepted: 2018-01-03T14:58:00Z. Cited on page 47.

[72] Joao Tiago Aparicio, João Salema de Sequeira, and Carlos J. Costa. Emotion analysis of Portuguese Political Parties Communication over the covid-19 Pandemic. In *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6, June 2021. ISSN: 2166-0727. Cited on page 47.

[73] Huong Nguyen Thi Xuan, Anh Cuong Le, and Le Minh Nguyen. Linguistic Features for Subjectivity Classification. In *2012 International Conference on Asian Language Processing*, pages 17–20, November 2012. Cited on page 55.

[74] Wei-Fan Chen, Khalid Al-Khatib, Benno Stein, and Henning Wachsmuth. Detecting Media Bias in News Articles using Gaussian Bias Distributions, October 2020. arXiv:2010.10649 [cs]. Cited on page 55.

[75] Gleice Carvalho de Lima Moreno, Marco P. M. de Souza, Nelson Hein, and Adriana Kroenke Hein. ALT: um software para an\'alise de legibilidade de textos em L\'ingua Portuguesa, August 2022. arXiv:2203.12135 [cs]. Cited on pages 58, 59, 60, 61, and 63.

[76] Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948. Place: US Publisher: American Psychological Association. Cited on page 59.

[77] Robert Gunning. *The Technique of Clear Writing*. McGraw-Hill, revised edition edition, January 1968. Cited on page 61.

[78] Meri Coleman and T. L. Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284, 1975. Place: US Publisher: American Psychological Association. Cited on page 61.

[79] E.A. Smith and R.J. Senter. *Automated Readability Index*. AMRL-TR. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, 1967. Cited on page 61.

[80] Ricardo Ribeiro Sanfins Moura. Automated Fake News detection using computational Forensic Linguistics. Master's thesis, Universidade do Porto, July 2021. Accepted: 2021-07-09. Cited on pages 63 and 80.

Appendix A

Data Collection Files

A.1 Brexit Configuration File

```
{
    "newspapers": [
      "https://www.theguardian.com"
    ],
    "date_range": {
      "start": "2019-01-01",
      "end": "2020-12-31"
    },
    "queries": ["Brexit", "European Union"],
    "output": "brexit.json"
}
```

Listing A.1: Example of a configuration file to search the terms “Brexit” and “European Union” in the The Guardian website between January 1st 2019 to December 31st 2020.

A.2 Opinion Articles Configuration File

```
{
    "newspapers": [
      "https://www.publico.pt",
      "https://www.dn.pt",
      "https://dnoticias.pt/",
      "https://observador.pt",
      "https://expresso.pt",
      "https://jornaleconomico.pt",
      "https://www.jornaldenegocios.pt"
    ],
    "date_range": {
      "start": "2018-01-01",
      "end": "2019-12-31"
    },
    "queries": ["opinião", "opiniao"],
    "output": "opinion.json"
}
```

Listing A.2: Example of a configuration file to search the terms related to “opinião” in a few Portuguese news websites between January 1st 2018 to December 31st 2019.

A.3 Reliable Articles Configuration File

```
{
    "newspapers": [
        "http://www.publico.pt",
        "http://www.jn.pt",
        "http://www.rtp.pt/",
        "http://expresso.pt",
        "http://jornaleconomico.pt"
    ],
    "date_range": {
        "start": "2010-01-01",
        "end": "2022-12-31"
    },
    "queries": [
        "tribunal",
        "estado",
        "governo",
        "política",
        "futebol",
        "desporto",
        "economia",
        "finanças",
        "saúde",
        "educação",
        "cultura",
        "ciência",
        "tecnologia",
        "ambiente",
        "justiça",
        "segurança",
        "religião",
        "música",
        "cinema",
        "literatura",
        "arte",
        "turismo",
        "lazer",
        "opinião",
        "vacina",
        "vacinação",
        "inflação",
        "taxas",
        "socialismo",
        "ps",
        "psd",
        "cds",
        "pcp",
        "be",
        "bloco de esquerda",
        "tap",
        "corrupção",
        "racismo",
        "racista",
        "xenofobia",
        "xenífobo",
        "fascismo",
        "hospital",
        "médicos",
        "enfermeiros",
        "professores",
        "professor",
        "professora",
        "alunos",
        "aluno",
        "aluna",
        "escola",
        "universidade",
        "universidades",
        "covid"
    ],
    "output": "reliability_only_reliable.json"
}
```

Listing A.3: Example of a configuration file to search diverse topics in a few Portuguese reliable news websites between January 1st 2010 to December 31st 2022.

A.4 Unreliable Articles Configuration File

```
{
    "newspapers": [
        "http://gazetapolitica.com/",
        "http://www.bombeiros24.pt/",
        "http://www.tuga.press",
        "http://www.direitapolitica.com",
        "http://noticiasviriato.pt",
        "http://portugalglorioso.blogspot.com",
        "http://www.lusopt.eu/",
        "http://inimigo.publico.pt/",
        "http://www.emdireto.pt"
    ],
    "date_range": {
        "start": "2010-01-01",
        "end": "2022-12-31"
    },
    "queries": [
        "tribunal",
        "estado",
        "governo",
        "política",
        "futebol",
        "desporto",
        "economia",
        "finanças",
        "saúde",
        "educação",
        "cultura",
        "ciência",
        "tecnologia",
        "ambiente",
        "justiça",
        "segurança",
        "religião",
        "música",
        "cinema",
        "literatura",
        "arte",
        "turismo",
        "lazer",
        "opinião",
        "vacina",
        "vacinação",
        "inflação",
        "taxas",
        "socialismo",
        "ps",
        "psd",
        "cds",
        "pcp",
        "be",
        "bloco de esquerda",
        "tap",
        "corrupção",
        "racismo",
        "racista",
        "xenofobia",
        "xenífobo",
        "fascismo",
        "hospital",
        "médicos",
        "enfermeiros",
        "professores",
        "professor",
        "professora",
        "alunos",
        "aluno",
        "aluna",
        "escola",
        "universidade",
        "universidades",
        "covid"
    ],
    "output": "unreliable.json"
}
```

Listing A.4: Example of a configuration file to search diverse topics in a few Portuguese fake and unreliable news websites between January 1st 2010 to December 31st 2022.

A.5 COVID-19 Articles Configuration File

```
{
    "newspapers": [
        "https://www.cmjornal.pt",
        "https://www.publico.pt",
        "https://www.dn.pt",
        "https://ionline.sapo.pt",
        "https://dnoticias.pt/",
        "https://www.sapo.pt",
        "https://www.rtp.pt/",
        "https://www.iol.pt/",
        "https://sol.sapo.pt/",
        "https://www.tvi24.iol.pt/",
        "https://noticias.sapo.pt/",
        "https://news.google.pt/",
        "https://observador.pt",
        "https://www.jn.pt",
        "https://www.tsf.pt/",
        "https://expresso.pt",
        "https://jornaleconomico.pt",
        "https://www.jornaldenegocios.pt",
        "https://caras.pt",
        "https://www.tv7dias.pt",
        "http://gazetapolitica.com",
        "http://www.bombeiros24.pt",
        "http://www.tuga.press",
        "http://tuga.press",
        "http://www.direitapolitica.com",
        "http://noticiasviriato.pt",
        "http://portugalglorioso.blogspot.com",
        "http://www.lusopt.eu",
        "http://lusopt.eu",
        "http://inimigo.publico.pt",
        "http://www.emdireto.pt",
        "http://jornaldiario.pt",
        "http://www.semanarioextra.com",
        "http://www.magazinelusa.com",
        "http://verdade.com.pt",
        "http://noticiario.com.pt",
        "http://voxpoptv.com",
        "http://noticiasdem3rda.com"
    ],
    "date_range": {
        "start": "2020-01-01",
        "end": "2022-12-31"
    },
    "queries": [
        "coronavirus",
        "coronavírus",
        "covid-19",
        "sars-cov-2",
        "pandemia",
        "estado de emergência",
        "estado de emergencia",
        "vacinação",
        "vacinacao",
        "distanciamento social",
        "máscara",
        "mascara",
        "quarentena"
    ],
    "output": "covid19.json"
}
```

Listing A.5: Example of a configuration file to search the terms related to COVID-19 in a few Portuguese news websites between January 1st 2020 to December 31st 2022.

A.6 Example Output File

The example is limited to 3 entries for simplification purposes, and the content is truncated. The full example can be accessed in the data collection tool GitHub page[1].

```
{
   "title":"Morais Sarmento já esperava que caso Portucale acabasse no tribunal",
   "content":"Morais Sarmento que, à data dos factos, era ministro da Presidência
       ↪ de Santana Lopes, respondeu hoje...",
   "date":"2011-06-09",
   "tags":"",
   "netloc":"publico.pt",
   "original_url":"http:\/\/www.publico.pt\/Sociedade\/morais-sarmento-ja-esperava-
       ↪ que-caso-portucale-acabasse-no-tribunal_1498220",
   "link_to_no_frame":"https:\/\/arquivo.pt\/noFrame\/replay\/20110609152700\/http
       ↪ :\/\/www.publico.pt\/Sociedade\/morais-sarmento-ja-esperava-que-caso-
       ↪ portucale-acabasse-no-tribunal_1498220"
}
{
   "title":"Escola, assembleia e executivo municipal no Tribunal da Boa-Hora",
   "content":"A passagem do tribunal para o município ocorre na sequência da extinç
       ↪ ão da Sociedade Frente Tejo, ontem aprovada em Conselho de Ministros...",
   "date":"2011-10-28",
   "tags":"",
   "netloc":"publico.pt",
   "original_url":"http:\/\/www.publico.pt\/Local\/tribunal-da-boahora-recebe-uma-
       ↪ escola-a-assembleia-municipal-e-o-executivo-1518680",
   "link_to_no_frame":"https:\/\/arquivo.pt\/noFrame\/replay\/20111028153921\/http
       ↪ :\/\/www.publico.pt\/Local\/tribunal-da-boahora-recebe-uma-escola-a-
       ↪ assembleia-municipal-e-o-executivo-1518680"
}
{
   "title":"Tribunal da Relação vai apreciar processo de magistrada alcoolizada",
   "content":"O processo relativo à procuradora Francisca Costa Santos que foi
       ↪ apanhada a conduzir alcoolizada e em contramão numa rua de Cascais, foi
       ↪ remetido como...",
   "date":"2014-10-21",
   "tags":"",
   "netloc":"publico.pt",
   "original_url":"http:\/\/publico.pt\/sociedade\/noticia\/tribunal-da-relacao-vai
       ↪ -apreciar-processo-de-magistrada-alcoolizada-1495159",
   "link_to_no_frame":"https:\/\/arquivo.pt\/noFrame\/replay\/20141021142543\/http
       ↪ :\/\/publico.pt\/sociedade\/noticia\/tribunal-da-relacao-vai-apreciar-
       ↪ processo-de-magistrada-alcoolizada-1495159"
}
```

Listing A.6: Example of an output file of the data collection tool.

[1] https://raw.githubusercontent.com/luist18/article-scraping/main/reliable_processed.json

Appendix B

Arquivo.pt API Requests Examples

B.1 Arquivo.pt CDX Server API Request

Query Query to the Arquivo.pt CDX server to get the preserved pages from a given URL, in the example below `sapo.pt` and filtered to the January 1st, 2023.

```
curl --location 'https://arquivo.pt/wayback/cdx?url=sapo.pt&from=20230101&to
    ↪ =20230101&output=json'
```

Listing B.1: Arquivo.pt CDX Server API curl command to get the preserved pages as a JSON from sapo.pt in January 1st, 2023.

Response JSON response of the query.

```
{"urlkey": "pt,sapo)/", "timestamp": "20230101081701", "url": "https://www.sapo.pt/
    ↪ ", "mime": "text/html", "status": "200", "digest": "
    ↪ XNMCQMTU4FPXIIAPQEX4ULF2WY5LWBJI", "length": "67955", "offset": "77438625",
    ↪ "filename": "fawp-20230101081231929-00002-6tlijxdr.warc.gz", "collection": "
    ↪ FAWP52", "source": "$root:FAWP52.cdxj", "source-coll": "$root"}
{"urlkey": "pt,sapo)/", "timestamp": "20230101081716", "url": "https://www.sapo.pt/
    ↪ ", "mime": "text/html", "status": "200", "digest": "
    ↪ NVSV5EBASS5GEUTI46FZW6QZQTUX633E", "length": "68683", "offset": "90834369",
    ↪ "filename": "fawp-20230101081231929-00002-6tlijxdr.warc.gz", "collection": "
    ↪ FAWP52", "source": "$root:FAWP52.cdxj", "source-coll": "$root"}
{"urlkey": "pt,sapo)/", "timestamp": "20230101175225", "url": "https://www.sapo.pt/
    ↪ ", "mime": "text/html", "status": "200", "digest": "
    ↪ C4IXZXVYUGQ3TQZYQQJET65TY3PD2BYR", "length": "68759", "offset": "156477", "
    ↪ filename": "WEB-20230101175219031-p97.arquivo.pt.warc.gz", "collection": "
    ↪ FAWP52", "source": "$root:FAWP52.cdxj", "source-coll": "$root"}
```

```
{"urlkey": "pt,sapo)/", "timestamp": "20230101175228", "url": "https://www.sapo.pt/
    ↪ ", "mime": "text/html", "status": "200", "digest": "
    ↪ LRM4THCVLHTP75WFCW7B2ZVPAGCBYHJ7", "length": "68761", "offset": "827260", "
    ↪ filename": "WEB-20230101175219031-p97.arquivo.pt.warc.gz", "collection": "
    ↪ FAWP52", "source": "$root:FAWP52.cdxj", "source-coll": "$root"}
{"urlkey": "pt,sapo)/", "timestamp": "20230101175237", "url": "https://sapo.pt/", "
    ↪ mime": "text/html", "status": "200", "digest": "
    ↪ EFLGIPW72FZGNKLCI5SMDDBJEGPDJ2N6", "length": "68758", "offset": "9203842", "
    ↪ filename": "WEB-20230101175219032-p97.arquivo.pt.warc.gz", "collection": "
    ↪ FAWP52", "source": "$root:FAWP52.cdxj", "source-coll": "$root"}
{"urlkey": "pt,sapo)/", "timestamp": "20230101175248", "url": "http://sapo.pt/", "
    ↪ mime": "text/html", "status": "302", "digest": "
    ↪ POPLDWWERZ2PUX2BRPCFNS2BB6ELQHMY", "length": "441", "offset": "47068782", "
    ↪ filename": "WEB-20230101175219020-p97.arquivo.pt.warc.gz", "collection": "
    ↪ FAWP52", "source": "$root:FAWP52.cdxj", "source-coll": "$root"}
{"urlkey": "pt,sapo)/", "timestamp": "20230101175846", "url": "http://www.sapo.pt/"
    ↪ , "mime": "text/html", "status": "302", "digest": "
    ↪ POPLDWWERZ2PUX2BRPCFNS2BB6ELQHMY", "length": "444", "offset": "65162429", "
    ↪ filename": "WEB-20230101175840676-p97.arquivo.pt.warc.gz", "collection": "
    ↪ FAWP52", "source": "$root:FAWP52.cdxj", "source-coll": "$root"}
```

Listing B.2: Query JSON response to get the preserved pages as a JSON from sapo.pt in January 1st, 2023.

B.2 Arquivo.pt Full-text search API Request

Query Query to the Arquivo.pt Full-text search API to get one article that matches the query “tribunal” in the Público website.

```
curl --location 'https://arquivo.pt/textsearch?type=html&siteSearch=publico.pt&q=
    ↪ tribunal&maxItems=1&dedupField=title&dedupValue=1'
```

Listing B.3: Arquivo.pt Full-text search API curl command to get one article matching “tribunal” from “publico.pt”.

Response JSON response of the query.

```
{
    "serviceName": "Arquivo.pt - the Portuguese web-archive",
    "linkToService": "https://arquivo.pt",
    "estimated_nr_results": 10743,
    "request_parameters": {
        "offset": 0,
        "dedupValue": 1,
        "type": [
            "html"
        ],
        "dedupField": "title",
        "q": "tribunal",
        "maxItems": 1,
        "siteSearch": [
            "publico.pt"
        ]
    },
    "response_items": [
        {
            "title": "Tribunal da Relação vai apreciar processo de magistrada
                ↪ alcoolizada - PÚBLICO",
            "originalURL": "http://publico.pt/sociedade/noticia/tribunal-da-relacao
                ↪ -vai-apreciar-processo-de-magistrada-alcoolizada-1495159",
            "linkToArchive": "https://arquivo.pt/wayback/20141021142543/http://
                ↪ publico.pt/sociedade/noticia/tribunal-da-relacao-vai-apreciar-
                ↪ processo-de-magistrada-alcoolizada-1495159",
            "tstamp": "20141021142543",
            "contentLength": 160170,
            "digest": "91cf4cb88f250331b7a134d290c1ba29",
            "mimeType": "text/html",
            "encoding": "UTF-8",
            "date": "1413901543",
            "linkToScreenshot": "https://arquivo.pt/screenshot?url=https%3A%2F%2
                ↪ Farquivo.pt%2FnoFrame%2Freplay%2F20141021142543%2Fhttp%3A%2F%2
                ↪ Fpublico.pt%2Fsociedade%2Fnoticia%2Ftribunal-da-relacao-vai-
                ↪ apreciar-processo-de-magistrada-alcoolizada-1495159",
            "linkToNoFrame": "https://arquivo.pt/noFrame/replay/20141021142543/http
                ↪ ://publico.pt/sociedade/noticia/tribunal-da-relacao-vai-apreciar
                ↪ -processo-de-magistrada-alcoolizada-1495159",
            "linkToExtractedText": "https://arquivo.pt/textextracted?m=http%3A%2F%2
                ↪ Fpublico.pt%2Fsociedade%2Fnoticia%2Ftribunal-da-relacao-vai-
                ↪ apreciar-processo-de-magistrada-alcoolizada-1495159%2
                ↪ F20141021142543",
            "linkToMetadata": "https://arquivo.pt/textsearch?metadata=http%3A%2F%2
                ↪ Fpublico.pt%2Fsociedade%2Fnoticia%2Ftribunal-da-relacao-vai-
                ↪ apreciar-processo-de-magistrada-alcoolizada-1495159%2
                ↪ F20141021142543",
            "linkToOriginalFile": "https://arquivo.pt/noFrame/replay/20141021142543
                ↪ id_/http://publico.pt/sociedade/noticia/tribunal-da-relacao-vai-
                ↪ apreciar-processo-de-magistrada-alcoolizada-1495159",
            "snippet": "<em>Tribunal</em> da Rela&ccedil;&atilde;o vai apreciar
                ↪ processo de magistrada alcoolizada - P&Uacute;BLICO Artigos
                ↪ seguintes Artigos anteriores T&oacute;picos Crime Lisboa <em>
                ↪ Tribunal</em> da Rela&ccedil;&atilde;o vai apreciar processo de
                ↪ magistrada alcoolizada 0 Partilhar no Facebook Partilhar no
                ↪ Twitter Partilhar no Google+ 36 <em>Tribunal</em> da Rela&ccedil
                ↪ ;&atilde;o vai<span class=\"ellipsis\"> ... </span>",
            "fileName": "IAH-20141021141309-61743-p13.arquivo.pt",
            "collection": "AWP16",
            "offset": 98019153
        }
    ]
}
```

Listing B.4: Arquivo.pt Full-text search API response to get one article matching “tribunal” from “publico.pt”.

Appendix C

OpenAI Chat Completion API

```
curl https://api.openai.com/v1/chat/completions \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer \$OPENAI_API_KEY" \
  -d '{
  "model": "gpt-3.5-turbo",
  "messages": [
    {
      "role": "user",
      "content": "Resume o seguinte texto para um texto fácil de interpretar para
          ↪ um leitor comum e capturando os principais pontos:\n\nSr. Presidente,
          ↪ Srs. Deputados, queria apenas deixar duas notas. A primeira é a de que
          ↪  consideramos absolutamente inaceitável que o Partido Socialista nas
          ↪ assembleias municipais onde esta questão foi discutida tenha votado a
          ↪ favor, tendo demonstrado um claro apoio aos trabalhadores das
          ↪ pedreiras, e nesta Assembleia da República dê o dito por não dito e
          ↪ vote contra um projecto de lei que é extremamente justo para os
          ↪ trabalhadores das pedreiras. Consideramos isso absolutamente inaceitá
          ↪ vel e de uma profunda contradição demagígica que importa aqui
          ↪ denunciar! A segunda é a de que a Sra. Deputada fala em prevenção. A
          ↪ petição dos trabalhadores foi discutida aqui, nesta Assembleia, há
          ↪ dois anos atrás. Que medidas de prevenção é que o Governo tomou sobre
          ↪ este sector? Zero, Sra. Deputada! Sra. Deputada, invocar a prevenção
          ↪ nos trabalhadores das pedreiras para chumbar um projecto de lei do PCP
          ↪  é imoral, tendo em conta que estes trabalhadores morrem a trabalhar!
          ↪ O Grupo Parlamentar do Partido Socialista é insensível a esta matéria!
          ↪  O Sr. Presidente: - Concluímos, assim, este ponto da ordem de
          ↪ trabalhos. Srs. Deputados, a Sra. Secretária vai dar conta da retirada
          ↪  de um diploma."
    },
  ],
  "temperature": 0.8,
  "max_tokens": 512,
  "top_p": 1,
  "frequency_penalty": 0,
  "presence_penalty": 0
}'
```

Listing C.1: OpenAI's Chat Completion API request example using data from the political stance dataset.

```
{
  "id": "chatcmpl-7a6ZwsyhdBV4c4kCjOt8RJdltinM9",
  "object": "chat.completion",
  "created": 1688839452,
  "model": "gpt-3.5-turbo-0613",
  "choices": [
    {
      "index": 0,
      "message": {
        {
            "role": "assistant",
            "content": "O discurso destaca a contradição do Partido Socialista por
                ↪ expressar apoio aos trabalhadores das pedreiras em assembleias
                ↪ municipais, mas votar contra um projeto de lei favorável a eles
                ↪ na Assembleia da República. Esse comportamento do Partido
                ↪ Socialista é criticado como inaceitável e hipícrita. A segunda
                ↪ crítica é sobre a inação do governo em relação à prevenção de
                ↪ acidentes trabalhistas no setor de pedreiras, apesar da discussã
                ↪ o anterior sobre essa preocupação. É citado como imoral usar a
                ↪ prevenção como motivo para rejeitar um projeto de lei,
                ↪ considerando que os trabalhadores estão morrendo em serviço. No
                ↪ final, é informado que a secretária vai anunciar a retirada de
                ↪ um diploma."
        }
      },
      "finish_reason": "stop"
    }
  ],
  "usage": {
    "prompt_tokens": 362,
    "completion_tokens": 135,
    "total_tokens": 497
  }
}
```

Listing C.2: OpenAI's Chat Completion API reply to the example above.