

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Portuguese Media Charts: Automated Media Assessment Using Large Language Models

João Carvalho



Mestrado em Engenharia e Ciência de Dados

Supervisor: Henrique Lopes Cardoso, PhD

Co-supervisor: Rui Carreira, PhD

July 1, 2025

Portuguese Media Charts: Automated Media Assessment Using Large Language Models

João Carvalho

Mestrado em Engenharia e Ciência de Dados

July 1, 2025

Abstract

Many times, media producers exhibit certain types of bias that are passed through to news consumers, potentially shaping their opinions and perception of the facts. This can have multiple negative effects in the form of polarized speech and general distrust on institutions and news producers themselves. There are several stages in the news production cycle at which media bias can be injected into the final product. In addition, there are also several types of bias. This dissertation approaches media bias that originates exclusively in the writing stage, mostly created through word choice and labeling. Also, the focus is on the European Portuguese landscape, which is very limited in terms of research in the field.

The main goal was to create an automated system that allows news consumers to directly frame news pieces according to four axes: political bias, reliability, objectivity, and readability. The classification methods are mainly focused on prompt-based techniques allied to open-weight large language models (LLMs) such as EuroLLM 9B and LLaMA 3.3 70B. Two specific techniques are explored: zero-shot prompting and a vector database technique tailored for the purpose. Additionally, a new prompt optimization procedure is presented based on the popular Particle Swarm Optimization (PSO) algorithm. Lastly, and exclusively for the readability axis, pre-established readability metrics are employed.

The results suggest that medium-sized LLMs are, in general, significantly more capable than small-sized ones, performing better on all tasks. However, the proposed vector database approach revealed to be a reliable way of improving the small model's performance. The 70B parameter model was able to reach, in some cases, an accuracy of approximately 90%, which is remarkable. Finally, the adaptation of PSO as a prompt optimizer revealed promising results, always contributing to an improved performance at the end of the process, showcasing clear optimization capabilities.

As the final product of this dissertation, the classification system is fully integrated with an interactive visualization tool in order to represent the Portuguese news landscape in interesting ways, providing the user a high level of freedom in terms of axis choice, number of dimensions (1D, 2D, or 3D plot), filtering, and a live-classification functionality. The exploration of the Portuguese case from multiple perspectives, the mix of techniques including state-of-the-art methodologies, a new prompt optimization methodology, as well as the development of an interactive visualization tool, are the main contributions of this work.

Keywords: Media bias, text classification, natural language processing, large language models, in-context learning, prompt-based methods

Resumo

Muitas vezes, os produtores de conteúdos mediáticos exibem certos tipos de enviesamento que são transmitidos aos consumidores de notícias, podendo moldar as suas opiniões e perceção dos factos. Isto pode ter múltiplos efeitos negativos, sob a forma de discurso polarizado e desconfiança generalizada nas instituições e nos próprios produtores de notícias. Existem várias etapas no ciclo de produção noticiosa nas quais o enviesamento mediático pode ser injetado no produto final. Além disso, existem também diversos tipos de enviesamento. Esta dissertação aborda o enviesamento mediático que se origina exclusivamente na fase de escrita, maioritariamente criado através da escolha de palavras e rotulagem. O foco recai sobre o panorama do Português Europeu, o qual apresenta uma escassez significativa de investigação nesta área.

O principal objetivo foi criar um sistema automatizado que permita aos consumidores de notícias enquadrar diretamente peças noticiosas de acordo com quatro eixos: viés político, fiabilidade, objetividade e legibilidade. Os métodos de classificação focam-se principalmente em métodos baseados em *prompts*, aliados a modelos de linguagem de grande escala (LLMs) abertos, como o EuroLLM 9B e o LLaMA 3.3 70B. São exploradas duas técnicas específicas: *zero-shot prompting* e uma técnica baseada em bases de dados vetoriais, concebida para este propósito. Adicionalmente, é apresentado um novo procedimento de otimização de prompts com base no algoritmo popular Particle Swarm Optimization (PSO). Por fim, e exclusivamente para o eixo da legibilidade, são utilizadas métricas de legibilidade previamente estabelecidas.

Os resultados sugerem que modelos de média dimensão são, em geral, significativamente mais capazes do que modelos de pequena dimensão, apresentando melhor desempenho em todas as tarefas. No entanto, a abordagem proposta com base em bases de dados vetoriais revelou-se uma forma fiável de melhorar o desempenho dos modelos mais pequenos. O modelo com 70 mil milhões de parâmetros foi capaz de alcançar, em certos casos, uma precisão próxima dos 90%, o que é notável. Finalmente, a adaptação do PSO como otimizador de *prompts* revelou resultados promissores, contribuindo sempre para um desempenho melhorado no final do processo, demonstrando claras capacidades de otimização.

Como produto final desta dissertação, o sistema de classificação está totalmente integrado com uma ferramenta interativa de visualização, de modo a representar o espaço noticioso Português de formas interessantes, proporcionando ao utilizador um elevado grau de liberdade em termos de escolha dos eixos, número de dimensões (gráfico 1D, 2D ou 3D), filtragem e funcionalidade de classificação em tempo real. A exploração do caso Português a partir de múltiplas perspetivas, a combinação de técnicas que incluem metodologias de ponta, uma nova metodologia para otimização de *prompts*, bem como o desenvolvimento de uma ferramenta interativa de visualização são as principais contribuições deste trabalho.

Keywords: Enviesamento mediático, classificação de texto, processamento de linguagem natural, modelos de linguagem de grande dimensão, *in-context learning*, métodos baseados em *prompts*

Acknowledgements

This dissertation represents the finish line of my Masters degree and two years of continuous learning and improvement. Throughout the development of this dissertation and the programme in general, I received the support of several people, to whom I am deeply grateful.

Firstly, I would like to thank my girlfriend, Francisca, who has helped me endure the demanding nature of this endeavor at all times with patience and words of encouragement during stressful phases. I am also grateful to my parents, whose invaluable support made all of it possible.

I want to express my gratitude to my supervisors, Henrique Lopes Cardoso and Rui Carreira, for their continuous guidance and expertise, which had a major impact in how this project has been shaped. Their genuine interest helped me stay on track and motivated throughout the whole process.

A word of gratitude also goes to every professor that I have crossed paths with during the degree, from whom I learned something new and incorporated it into this research, directly or indirectly. Finally, I want to thank the colleagues with whom I closely worked on projects and assignments, who also contributed to my success on this journey.

Thank you all for the invaluable contributions.

João

*“Success is the sum of small efforts,
repeated day in and day out.”*

Robert Collier

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Problem and Objectives	3
1.3	Structure	5
2	Literature Review	7
2.1	Media Bias	7
2.1.1	Introduction and Definitions	7
2.1.2	Linguistic Bias	9
2.1.3	Addressing Media Bias	10
2.1.4	Existing Tools	11
2.2	Classification Methods	15
2.2.1	Traditional NLP	16
2.2.2	Traditional Machine Learning Methods	18
2.2.3	Deep Learning Methods	19
2.2.4	Prompting Methods and Optimization	24
2.2.5	In-Context Learning	26
2.3	Research Gaps	29
3	Research Methodology	31
3.1	Framing Axes	31
3.2	Datasets	32
3.2.1	Political Bias Dataset	32
3.2.2	Reliability Dataset	33
3.2.3	Objectivity Dataset	34
3.3	News Articles Classification	35
3.3.1	Zero-shot Classification	36
3.3.2	Vector Databases	38
3.3.3	Readability Assessment	41
3.4	Prompt Optimization	43
3.4.1	Particle Swarm Optimization	43
3.4.2	Particle Swarm Prompt Optimization	44
4	Experimental Evaluation	47
4.1	Results	47
4.2	Particle Swarm Prompt Optimization	49
4.3	Answering the Research Questions	51
4.3.1	Research Question 1	51

4.3.2	Research Question 2	51
4.4	Indicator Analysis	52
4.4.1	General Statistics	52
4.4.2	Indicator Vector Space	53
4.5	Readability Assessment	54
4.6	Axes Relationships	56
4.7	Real World Takeaways	58
5	Interactive Visualization Tool	61
6	Conclusion	65
6.1	Summary	65
6.2	Contributions	66
6.3	Future Work	67
A	Political Bias Dataset	69
B	Prompts Utilized	71
B.1	Zero-shot Prompts	71
B.1.1	Political Bias	71
B.1.2	Reliability	72
B.2	Vector Database Prompts	72
B.2.1	Political Bias	72
B.2.2	Reliability	74
B.2.3	Objectivity	76
B.3	Particle Swarm Prompt Optimization	78
C	Visualization Tool Views	81
	References	85

List of Figures

2.1	Ad Fontes Media bias chart (Ad Fontes Media, 2025)	12
2.2	AllSides bias chart (AllSides, 2024)	13
2.3	Biasly bias chart (Biasly, 2025)	14
2.4	LSTM unit architecture	20
2.5	BERT for text classification (from Devlin et al. (2018))	22
3.1	Vector databases classification pipeline	38
4.1	Optimization trajectory on political bias. Each line represents one particle (prompt).	50
4.2	Indicator embeddings in 2 dimensions. The blue color is assigned to the classes “left”, “reliable”, and “subjective”, in the respective order of axes; red is assigned to the remaining classes.	54
4.3	Correlation matrix for readability metrics	55
4.4	Readability scores distributions for every metric	56
4.5	Relation between reliability and objectivity	57
4.6	Pearson correlations between axes	58
5.1	Tool UI with 2D plot using “Group By” functionality	62
5.2	Hover data shown for one article	63
5.3	Live classification functionality input	63
C.1	1D plot for political bias filtered for two sources	81
C.2	Unfiltered 2D plot for reliability and objectivity	82
C.3	Unfiltered 3D plot for reliability, political bias, and objectivity	82
C.4	1D plot for readability with source filters, grouping functionaliy, and live classifi- cation through URL	83

List of Tables

1.1	Example of media bias in news headlines identified by AllSides, a website dedicated to identifying political bias in news articles.	2
3.1	Political bias dataset source and class distribution	33
3.2	Fake News PT source distribution	34
3.3	MINT's class distribution	35
3.4	Hard news and opinion articles source distribution	35
4.1	Performance comparison across models for political bias, reliability, and objectivity. "ZS" refers to the zero-shot methodology; "VD" refers to the vector database approach. The F1-score metric presented is the weighted-averaged version. . . .	47
4.2	Performance comparison for political bias, reliability, and objectivity. "ZS" refers to the zero-shot methodology; "PSPO" refers to the prompt-optimized alternative. The F1-score metric presented is the weighted-averaged version.	50
4.3	Number of unique generated indicators in each vector database	52
4.4	Average scores per source on each axis	59
A.1	Authors considered in the political bias dataset	70

Abbreviations

DL	Deep Learning
GBT	Gradient Boosted Trees
GRU	Gated Recurrent Unit
k-NN	k-Nearest Neighbors
ICL	In-context Learning
LLM	Large Language Model
LR	Logistic Regression
LSTM	Long-Short Term Memory
ML	Machine Learning
NB	Naive Bayes
NLP	Natural Language Processing
NN	Neural Network
POS	Part-of-Speech
RF	Random Forest
RNN	Recurrent Neural Network
SVM	Support Vector Machines
XGB	Extreme Gradient Boosting

Chapter 1

Introduction

The media play a crucial role in shaping public opinion and influencing societal discourse (Hamborg et al., 2019). With this power comes the responsibility of adhering to journalistic principles of factuality, objectivity, and neutrality. However, media bias remains a persistent issue, often manifesting itself through framing, selective fact presentation, and sensationalism (Spinde et al., 2023). While some level of subjectivity is perhaps inevitable, unrecognized biases can distort public perception and erode trust in mainstream news sources.

This dissertation addresses the challenge of media assessment and bias detection by exploring automated methods to classify and analyze news content. By leveraging advanced natural language processing techniques, the research aims to provide a structured and explainable approach to identifying bias, with a specific focus on Portuguese media, a largely underexplored area in this field.

1.1 Motivation

Media outlets produce, on a daily basis, informational and news content on a variety of different topics that are consumed by the masses. The Portuguese Journalists Union published a deontological code¹ regarding the creation of news content, establishing basic journalistic principles such as factuality, objectivity, combat against sensationalism, accountability, and neutrality, among others. One of the main reasons for the cruciality of these principles is the power of media outlets to shape public opinion and the responsibility they bear while doing so (Dallmann et al., 2015; Hamborg et al., 2019; Spinde et al., 2023). Many media outlets deviate from some of the principles cited above, exhibiting considerable bias towards certain views or aspects of societal discourse (Wolton, 2019; Spinde et al., 2023).

According to Hamborg (2022), news is inherently not neutral, as it contextualizes facts and assesses events, which can lead to framing and biased reporting. However, the real issue arises

¹<https://jornalistas.eu/codigo-deontologico/>

when consumers are unaware of these biases, which can distort public opinion and policy decisions (Wolton, 2019). Media bias manifests itself through various mechanisms, including word choice, fact selection, and imagery, all of which influence how events are perceived. The high effort required for critical assessment, such as comparing multiple sources, often deters news consumers from challenging biases during routine consumption (Hamborg, 2022).

A concrete example of what media bias might look like is some of the coverage given to Mark Zuckerberg’s interview on Joe Rogan’s podcast “The Joe Rogan Experience”. This interview was conducted following Meta’s decision to end the social network giant’s fact-checking policy and replace it with a mechanism called Community Notes. Headlines about this interview are an excellent example of how the same event can be framed in different ways; if these different headlines are put beside each other, the different feelings conveyed purely by the use of words become clear.

Table 1.1: Example of media bias in news headlines identified by AllSides, a website dedicated to identifying political bias in news articles.

Media Outlet	News Headline	AllSides Classification
New York Post	<i>Biden officials ‘screamed’ and ‘cursed’ at Meta execs to take down vaccine posts, Mark Zuckerberg tells Joe Rogan</i> ^a	Right-wing biased
Newsweek	<i>Mark Zuckerberg Unloads on Biden White House in Joe Rogan Interview</i> ^b	Center
Axios	<i>Zuckerberg calls for “repopulation” of “cultural elite class”</i> ^c	Left-wing biased

^a New York Post.

^b Newsweek.

^c Axios.

Table 1.1 shows three headlines for the same event, produced by different outlets. The New York Post’s title transcribes some of the words said in the interview to gather the attention of the reader and frame Biden in a negative way right from the start. Additionally, the title specifically mentions the topic of vaccine posts, which is controversial in itself (Rao and Andrade, 2011). A more moderate title is the one from Newsweek, which does not specify any topic and goes straight to the point. Lastly, Axios’ piece gives little context in the headline, but within the main body, the article describes Zuckerberg’s words as “common Trumpism”, negatively framing both the whole situation and Donald Trump himself. The respective AllSides classification is shown in the table as well. A comparative analysis like this is a clear way to identify media bias; however, it may not be easy for readers to read multiple articles on the same matter (Spinde et al., 2023), hence the importance of automated methods.

According to a yearly poll performed in the United States (Megan Brennan, 2024), trust in

the media is at a historic low. Americans' trust in the media has decreased to a point where, in 2024, only 31% of respondents said they believed what they were presented by mainstream media outlets. This is down from 55% in 1999. On the other hand, during the same period, the percentage that showed no trust at all went from 11% to 36%. Without inferring any causal relationship here, the importance of maintaining a neutral and objective stance when informing the public can still be highlighted, even for the sake of the media outlets themselves. Without trust in the mainstream media, people start looking for alternative sources that also have their own weaknesses and biases, fomenting phenomena like filter bubbles and echo chambers, contributing to the polarizing narratives that are part of the discourse of many political and social matters these days (Spinde et al., 2022a).

There is vast research addressing this topic from multiple perspectives (Spinde et al., 2023), and even some recent tools that help identify media bias. Examples of these tools are the media chart developed by Ad Fontes Media (Ad Fontes Media, 2025), an American corporation that focuses its business around news classification, and similar platforms offered by AllSides (AllSides, 2024) or Biasly (Biasly, 2025), with the last two mainly focused on political news and partisan bias.

Highlighting media bias instances has positive implications and can mitigate the effects of such biases (Baumer et al., 2015). It can also enable journalists and publishers to assess their work objectively (Dallmann et al., 2015). While media literacy is a potential solution to empower consumers to critically assess news, its high effort requirements make it impractical for daily news consumption, leaving the public vulnerable to the influence of media bias.

Although there is indeed a large amount of scientific research on media bias (Spinde et al., 2023), most of this research is done for the English language, and in particular, it covers the American news media landscape. This leaves a gap in several different languages – a major motivation for the completion of this dissertation, which provides additional research covering the underexplored European Portuguese media landscape.

1.2 Problem and Objectives

Media outlets are often biased in some way (Spinde et al., 2023; Wolton, 2019). As shown by Rodrigo-Ginés et al. (2024), media bias can take multiple forms, each with its own nuances and implications. In this dissertation, several forms are taken into consideration. Since manually classifying a news article according to a certain type of bias, let alone several, is very complex and time-consuming, the main objective set for this dissertation was to develop an effective system to automate this process. The goal is not only to automate the detection but also to allow for some level of explainability by pointing out where the origins of bias are within the text. This is a very relevant factor that is lacking in many automated methods (Bhattarai, 2024).

There are many ways in which the framing of news articles into several axes can be done. Some of these possibilities are described in detail in Section 2.2. In this dissertation, the primary way was to take advantage of the deep linguistic understanding capabilities of Large Language Models

(LLMs) and In-context Learning (Dong et al., 2022), a family of methods often referred to as prompt-based techniques. This is a paradigm that emerged relatively recently with the appearance of modern-day decoder LLMs, such as GPT-3 (Brown et al., 2020). These models are currently trained on such large amounts of data, with the set goal of modeling language, that they end up being capable of performing remarkably well in several natural language processing tasks, just by prompting them to do so. What is assessed in this dissertation is whether the performance can be maintained when the method is applied to a nuanced text classification task, such as the detection of bias in news articles, and by using small- and medium-scale LLMs that can be run in consumer-grade hardware. From this goal, the first research question arises:

1. Are small and medium-scale LLMs capable of successfully framing news articles into media bias-related topics?

This is the most general question that encompasses the main goal of this project. It covers the most prominent part of the methodology regarding the use of prompting techniques and highlights the specific use of models that run on consumer-grade hardware.

One important factor of current prompt-based techniques, when paired with instruction-tuned models such as LLaMA (Touvron et al., 2023), for instance, is the act itself of prompting the model. It has been shown that prompting can be optimized towards better end results (Liu et al., 2023) at the expense of, for instance, providing some examples within the prompt itself, so that more context is given about the intended task – a technique referred to as Few-shot Learning (Lin et al., 2024). The second goal of the dissertation was to develop a novel way of optimizing prompts inspired by a popular physics-based optimization algorithm, therefore defining the second research question as follows:

2. Does Particle Swarm Optimization perform well as a prompt optimizer?

Guo et al. (2023) built a prompt optimization framework based on Evolutionary Algorithms and denoted promising results. Building on this, the objective was to assess how much performance could be improved by adapting a popular optimization algorithm, such as Particle Swarm Optimization, as a prompt optimizer for the task of media bias detection. The adaptation is done so that it works with natural language prompts and with the optimization operator being an LLM itself.

The final objective of this project was to develop an interactive visualization tool that seamlessly integrated the classification system, and allowed for a pleasant and flexible user experience. No research question was materialized related with this goal.

This dissertation poses as a valuable contribution on multiple fronts. It advances various aspects of media assessment, and explores the use of novel techniques, such as leveraging LLMs and their powerful linguistic understanding for text classification, using an in-context learning approach. It also contributes to research on prompt optimization methods. Furthermore, it includes the development of an interactive visualization tool that serves as a front-end for the whole system.

Lastly, this work also supports research covering Portuguese media, which is a big challenge in itself.

1.3 Structure

The remainder of this document is structured as follows. Chapter 2 presents a comprehensive literature review that addresses media bias from multiple perspectives, its origins, driving factors, and ways of tackling it from a detection standpoint. Additionally, in the second half of the chapter, the main text classification methods are explored and contextualized in a media assessment setting. Chapter 3 maps out the development process of the dissertation by presenting an overview of the axes in which news articles are framed in this context, the datasets used, the two classification and framing methods developed, and finally the prompt optimization technique created in the context of this dissertation. Chapter 4 presents the results of each method and the performance of the optimization algorithm, followed by a direct response to the research questions posed, and ending with a deeper analysis of the methodologies employed and their results. Chapter 5 is exclusively dedicated to the development process and functionalities of the visualization tool. Lastly, Chapter 6 concludes by presenting the contributions of the dissertation and proposals for future research.

Chapter 2

Literature Review

This literature review offers a complete overview of the existing knowledge relevant to the research topic and existing methodologies. The chapter's first section concentrates on media bias research from a generic point of view and the different ways media bias can arise. In the second section, the goal is to explore past and present techniques that can be employed for text classification and, in particular, for media bias detection on news articles. The chapter ends by highlighting the significance of this research and articulates its unique contributions.

2.1 Media Bias

This section addresses media bias research from a generalist point of view. It starts by presenting important definitions, terminology, and different ways in which media bias can originate and manifest. Then, it focuses on the specific type of bias of interest for this dissertation: linguistic bias. After this, it goes over the different strategies to tackle media bias and, lastly, offers a comprehensive overview of existing tools that serve as inspiration for one of the main objectives – the interactive visualization tool.

2.1.1 Introduction and Definitions

Technology has revolutionized how individuals access knowledge, offering a number of different ways to engage with information and societal issues. In particular, online news has gained significant importance, surpassing traditional media formats such as TV, radio, and print media (Dallmann et al., 2015). However, the quality of this engagement relies heavily on the availability of unbiased sources of information. This is an essential factor for informed decision-making and the formation of well-rounded opinions. Media bias, described by Spinde et al. (2023) as an intentional or unintentional expression of tendency in favor of a perspective, ideology, or result, undermines this ideal by introducing systematic slants in how events are reported. Rooted in factors such as ownership, funding, and ideological influences (University of Michigan), bias in

media coverage can distort public perception, shaping the general public's opinions and decisions (Dallmann et al., 2015; Hamborg et al., 2019; Spinde et al., 2023).

Media bias manifests through various mechanisms, including selective word choice (Grefenstette et al., 2004), the credibility ascribed to sources (Gentzkow and Shapiro, 2006), and the visual prominence given to articles in the form of size and placement (Bucher and Schumacher, 2006). An important factor that might be considered is the concentration of the news provider market, where a small number of corporations exert significant control over the media landscape. This concentration amplifies the risk of intentional bias, as editorial decisions driven by financial or ideological agendas would be more impactful. In Portugal, for example, the four largest newspapers for informational news (Correio da Manhã, Expresso, Jornal de Notícias, and Público) corresponded to over 90%¹ of the paid circulation (considering both paper and online circulation) from the period of January to March 2025, the last quarter with available data as of writing. Furthermore, some of the owners of these newspapers are media conglomerates with a presence in other newspaper types and magazines, TV stations, and radio, increasing their overall reach into every form of mainstream media.

Some of the first research pieces on media bias date back several decades, to at least 1950, when White (1950) analyzed the selection process of a newspaper. In 1975, Williams (2006) described media bias as something that has to be intentional and sustained, therefore distinguishing intentional media bias from unintentional influences. Different forms and definitions of media bias exist, often dependent on specific research contexts. For instance, Mullainathan and Shleifer (2005) proposed ideological bias, which is characterized by promoting specific opinions and views, and spin bias, which the authors defined as framing facts in a way that creates memorable stories. Another common framework is from D'Alessio and Allen (2006), which covers the impact of media bias during US Presidential elections, and divides media bias into coverage bias (visibility of topics or entities relative to each other), gatekeeping bias (cherry-picking of stories or omission of others to convey a certain view), and statement bias (tone and coverage presentation of information).

Additionally, an important perspective to consider is the stage at which media bias arises in the news production process. Between reality and the actual news about reality, there are several steps where bias can be added to the final product. On this front, Park et al. (2009) described the news creation process as a sequence of three phases: the gathering phase, the writing phase, and the editing phase. During the gathering stage, four elements could potentially turn into sources of media bias: commission and omission represent decisions about which details to include or exclude; story selection is about journalists choosing which events to report, often influenced by relevance, sensationalism, or target audience preferences; and lastly, source selection happens when journalists pick sources such as press releases or studies, shaping the narrative based on the sources used. In the writing phase, three other sources appear: labeling, by stigmatizing individuals, groups, or other entities as positive or negative; word choice, perhaps one of the most subtle and yet prominent and direct ways of biasing a news article, using specific terminology to frame events and

¹ ACPT - retrieved on May 17th, 2025

resourcing to words and expressions with strong meaning, conveying a message; and tone, which is also set by the choice of specific words. Before reaching the end of the cycle, according to [Park et al. \(2009\)](#), news articles still go through the editing phase. At this stage, common origins of bias are related to the visual appearance of the articles; for example, the placement of the article within the overall context of the page or website, or the selection of photos, which is an easy way of showing facts in a distorted or out-of-context way. [Matthes et al. \(2023\)](#) and [Peng \(2018\)](#) also proposed similar ideas, essentially stating that media bias doesn't come only in text format, but also via pictures or the overall layout of the news. [Peng \(2018\)](#) used computer vision techniques to analyze thousands of images from news reporting on the 2016 US presidential election, and found that Hillary Clinton was often portrayed more favorably and associated with happiness, while Donald Trump was usually represented as a less positive, angry, and more dominant figure.

The motivations driving such factors of bias are different in nature ([Hamborg, 2022](#)). Internal motivations such as the views and biases of the journalists themselves or the media outlet's management ([Groseclose and Milyo, 2005](#)) are one example. Another type of motivation might come from external factors such as the influence of consumer preferences, a motivation stronger than the previous one, according to [Gentzkow and Shapiro \(2010\)](#). [Atkins \(2016\)](#) also explained media bias originates to lure larger audiences and, at the same time, is cheaper to produce than objective, unbiased news. Pressure from advertisers, which are a source of income for news producers, or even the government itself, also appear as major reasons for news producers to create biased content ([Besley and Prat, 2006](#)).

2.1.2 Linguistic Bias

Although the gathering and editing stages of news production are important for the full framing of media bias, this dissertation's primary research field lies within Natural Language Processing (NLP). With that said, the most relevant stage of news production introduced in the previous section is the writing phase. In this section, different research is presented focusing on media bias solely originating at this particular stage. This means this research does not assess the degree to which a news article is biased in source selection or story selection (gathering phase), and not even with image selection or article placement (editing phase). To make such an assessment, more than the articles themselves would be needed. On the other hand, to evaluate wording bias or tone bias, the article's text is the only requirement.

Bias by word choice is a very relevant source of bias ([Spinde et al., 2021c](#)), and sometimes, researchers refer to it with different terminology, such as linguistic bias ([Recasens et al., 2013](#); [Beukeboom and Burgers, 2017](#)), or lexical bias ([Fan et al., 2019](#)). This refers to systematic usage of certain words, phrases, and terminology that induces a particular view, for or against a particular person, group, or entity, reflecting stereotypes or preconceived notions ([Beukeboom and Burgers, 2017](#)).

[Papacharissi and de Fatima Oliveira \(2008\)](#) and [Feick et al. \(2021\)](#) studied the impact of word choice on terrorism-related news. The first one argued that different media outlets use a more dramatic tone, for example, using hyperboles and metaphors, others a less dramatic tone, and others

focus solely on the factual reporting of events. Feick et al. (2021), on the other hand, concluded that the tone of media coverage significantly influences public perceptions, with sensationalist and negative language heightening perceived threats. Also, sensationalist reporting can fuel unjustified fears, distort public perceptions, and even have economic consequences such as reduced tourism and exaggerated narratives of unrest. Unfortunately, according to Feick et al. (2021), media outlets seem to often prioritize sensationalism for attention and commercial success, despite its broader societal impact.

Druckman and Parkin (2005) explored how media slants in tone and word choice affect electoral outcomes. The research found that editorial biases significantly influence voter perceptions of candidate image traits and even question the media's power within a democratic society. Stefánsson (2014) explored evaluative labeling in news, revealing that labels often carry connotations that skew perceptions of individuals or groups. For instance, labeling political parties as “radical” or “progressive” can manipulate audience attitudes. Groeling (2013) highlighted the disparities in tone between news outlets, showing that partisan outlets provide systematically different valences in coverage, affecting electoral outcomes. The author concluded by noting that in the 21st century, partisan journalism in the USA is very much a reality, resembling much more the times of the 19th century, other than the 20th century, the “golden age of objective press”.

As denoted, journalists can use certain words and terminology to frame a given story in a positive or negative context, without changing the overall narrative or actively lying. A concrete showing of this would be some of the coverage given to a fairly recent event, at the time of writing, in which an explosion originated from a Tesla vehicle. Giving a specific example from the Portuguese media, the title goes as follows: “Veículo da Tesla explode e faz um morto à porta de hotel de Donald Trump em Las Vegas”². What the title fails to make clear is that it wasn't exactly the vehicle that exploded, but the fireworks and gas bottles inside it. The article mentions this at the end, but intentionally or not, in the title, as well as in the introduction, it gives the impression that the vehicle itself caught on fire and exploded.

2.1.3 Addressing Media Bias

Media bias can be addressed during news production or after the news is produced. Efforts during production aim to prevent bias by fostering objective reporting or offering formats like press reviews, opinion articles, or commentaries that freely expose biases. However, these approaches are seen as impractical due to the subjective nature of news selection, the challenge of defining objectivity itself, and the structural bias inherent in news coverage (Park et al., 2009). Financial incentives further complicate objectivity, as publishers often cater to audience ideologies (Gentzkow and Shapiro, 2010).

Given these challenges, post-production methods offer a more feasible way of addressing media bias. These fall into three categories: bias analysis, bias correction, and bias communication (Hamborg, 2022). Bias analysis involves identifying and analyzing biases, either manually or

²Unclear headline from Observador

through automated systems. While manual methods are effective and able to capture the nuanced nature of biased news, they are resource-intensive, whereas automated methods are scalable but sometimes end up with superficial results, struggling to capture this nuance and generalizing poorly outside of their domain (Lin et al., 2024). Bias correction seeks to rectify biases by replacing biased statements with neutral ones; for instance, flipping the ideological slant of headlines is one tested application by Chen et al. (2018). According to Hamborg (2022), bias correction methods have shown poor success, often producing incoherent results and showing the difficulty and subjectivity of the task. In Section 2.2.3.2 there is some literature proving otherwise (Pryzant et al., 2019; Raza et al., 2024). Lastly, bias communication aims to directly inform consumers about biases in news coverage. Techniques like bias-sensitive visualizations and tools like AllSides or live web browser tools (Afonso and Rosas, 2024) present diverse perspectives, helping users recognize biases and form more balanced views. Research has shown that bias communication can enhance informed decision-making and foster greater perspective diversity (Baumer et al., 2017); however, its effectiveness is limited by the challenges of identifying biases accurately and the reliance on resource-intensive methods.

To conclude, post-production approaches, and in particular bias analysis and bias communication, are the most promising and suitable for tackling media bias in the context of this dissertation.

2.1.4 Existing Tools

As mentioned in Section 1.1, there are some tools and platforms that are also part of the motivation behind this dissertation. In this section, a closer look is given to what exactly these tools offer, and what kind of functionalities served as inspiration to the tool that ultimately developed as the culmination of all the work produced in this research.

2.1.4.1 Ad Fontes Media

Ad Fontes Media is an American for-profit company based in the state of Colorado and founded in 2018. Their main stated source of value is to thoroughly analyze media of various formats with the objective of highlighting misleading, inaccurate, and highly polarizing media content. The company sells products around this concept to advertisers, publishers, educators, and even individuals. Within the several products, there is one that is especially interesting considering the scope and goals of this dissertation: the Interactive Media Bias Chart.

The chart contains two axes, one relative to News Value and Reliability, and the other to Political Bias, both measured quantitatively as well as qualitatively. The first one ranges in value from 0 to a maximum of 64 and is also characterized by 8 categories that go from “Contains Inaccurate / Fabricated Info”, the least reliable type, to “Original Fact Reporting, High Effort”, the most reliable. As far as the second axis goes, it ranges from -42 to 42, in a way that represents a one-dimensional political spectrum that goes from “Most Extreme Left” to “Most Extreme Right”, with everything else in between.

Media. It focuses exclusively on political bias and written news articles, framing each article into five possible classes, in the following order: “Left”, “Lean left”, “Center”, “Lean right”, and “Right”.

On the front page of the official website, one can access a wide variety of already-classified news about recent events and political discourse. One very convenient functionality is the showing of news from multiple media outlets reporting on the same event side-by-side, accompanied by the AllSides classification. This gives a very clear view of how the authors biased the headline or story in a specific way, according to their own beliefs and motivations. An example of this is showcased in Table 1.1.



Figure 2.2: AllSides bias chart (AllSides, 2024)

Figure 2.2 shows the media bias chart offered by AllSides, which, contrary to the one from Ad Fontes Media, is just one-dimensional. It presents a variety of media outlets according to their overall bias, and by clicking on any of them, a dedicated page is loaded, providing more details on how and why the score was given, as well as a few examples of news to show.

Regarding methodology, AllSides employs a mix of inputs coming from a multi-partisan editorial review done by trained experts, as well as blind bias surveys conducted with the public. Therefore, it includes not only the opinion of a limited group of analysts but also of a wider group of average Americans.

2.1.4.3 Biasly

Biasly, founded in 2017, addresses media bias in political news and classifies them according to their political bias, similarly to the platforms presented before. At the core of its work is the Bias Meter, a machine-learning-based tool that analyzes hundreds of articles on a daily basis to assess their political bias. Similarly to Ad Fontes Media, Biasly doesn't go much further into the technicalities of the tool, but it highlights the use of deep neural networks and NLP techniques to identify political sentiment and evaluate tone and language. Complementing this technology, a diverse team of analysts reviews articles using a standardized scoring system, assigning a bias score on a scale from -100 (Very Liberal) to +100 (Very Conservative), with 0 being the neutral point (Center). Scores from liberal, moderate, and conservative analysts are averaged for accuracy.

On its homepage, the website offers a high variety of news articles already classified, as well as a "Side by Side" area that shows examples of the same story reported by different media companies and with different bias classifications, very similar to what was previously described as a feature available in the AllSides website. Additionally, Biasly offers a Google Chrome extension that enables users to obtain bias ratings for news articles on a live basis, highlighting the importance and capabilities of automated systems in terms of scalability, speed, and convenience.

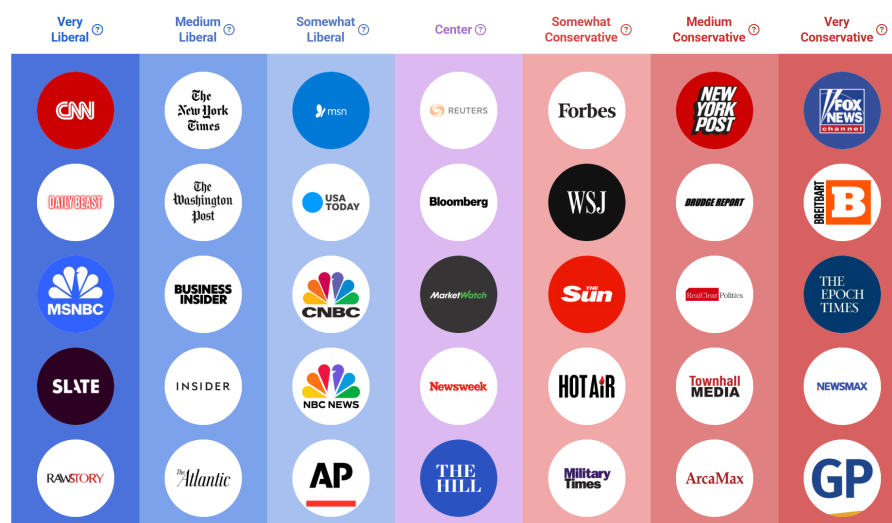


Figure 2.3: Biasly bias chart (Biasly, 2025)

Like the previously presented platforms, Biasly also offers a visually-friendly media bias chart that looks and works very similarly to the one offered by AllSides. Figure 2.3 shows Biasly's take on the media bias chart, where each media outlet is classified according to its overall score and assigned to a specific class in the political spectrum. Clicking on any of the outlet logos takes the user to a separate page showing some context and explanation of the classification itself, including the exact value on the determined scale.

2.1.4.4 Solutions in Portuguese

Every tool presented in Section 2.1.4 at this point has a particular characteristic in common that is important to highlight in the context of this dissertation: they target news written in English, particularly the American landscape. In European Portuguese, there is no tool or resource, manual or automated, that compares to the level of depth provided by the tools presented before. In spite of this, there is some work that is worth considering during the planning phase of this dissertation.

Firstly, some news providers have fact-checking platforms that serve as a tool for verifying the accuracy of information after it has been disseminated to the public. The main technical difference between most of these fact-checking platforms and the ones presented before is that these operate by cross-referencing claims with reliable sources and consulting experts, whereas political bias charts usually rely solely on the linguistics of the text for the classifications. This is an integral difference because, for automated systems, the text itself is often the only thing available, and an automated fact-checking tool using an external knowledge base would demand a lot more from a development and maintenance standpoint. With that said, examples of these Portuguese fact-checking platforms would be *Polígrafo* from *Sapo*⁴, *Fact Check* from *Observador*⁵, and *Prova dos Factos* from *Público*⁶. The three platforms focus on combating misinformation by manually checking the validity of claims and rumors, but the vast majority of what is analyzed are either controversial social media posts from average people, or claims done publicly by some kind of political figure, with very little relation to actually looking into news reported by the media in general. This is a very big distinguishing factor between common fact-checkers and what is targeted in this dissertation.

A better comparison would be the contribution of [Tavares \(2023\)](#), who created an automated system mostly based on deep learning methods to frame news articles into several axes. With a similar inspiration in the tools already introduced in previous sections, [Tavares \(2023\)](#) also developed a prototype of a bias chart most similar to the one offered by Ad Fontes Media, since it allowed for a 2-dimensional plot. On the other hand, [Afonso and Rosas \(2024\)](#) developed a somewhat different but interesting tool that also fits within the news classification topic in the form of a mobile application and Google Chrome extension for the detection of fake news, similar to some of the services provided by Biasly. These are possibly the two best examples of automated news article classification systems, with an integrated tool specifically built with the European Portuguese landscape in mind. The methods employed by both are further explored in Section 2.2 of this literature review.

2.2 Classification Methods

As mentioned before, when showcasing the motivation and problem to be addressed, the development and research of this dissertation are framed within the field of NLP. More specifically, the

⁴Polígrafo

⁵Observador Fact Check

⁶Prova dos Factos

task can be described as text classification, one of the most researched tasks in the field. Because of this, there is a large variety of methods to perform this type of classification.

Text classification is very much used to extract sentiment from text, whether it be from product reviews (Haque et al., 2018; Sinha, 2024), social media interactions on Facebook or Twitter (now X) (Go et al., 2009; Yue et al., 2019), or even for company statements/policy statements (Mućko, 2021; Gorodnichenko et al., 2023). The extracted sentiment from some of these examples could be used for things like recommender systems (Chen et al., 2015) or stock price predictors (Cakra and Trisedya, 2015). Other examples of tasks based on text classification could be spam detection (Crawford et al., 2015), and document tagging, that is, assigning relevant tags or keywords to documents based on their content (Chen et al., 2017).

This section reviews the available methodologies for text classification in detail, machine learning-based or not, by presenting the foundational research that introduced them and showcasing practical or theoretical applications where these methodologies proved useful, specifically targeting text classification within a media bias context. With this, parallels can be drawn, and many specificities of the applications themselves also make sense within the context of this dissertation.

2.2.1 Traditional NLP

The origins of traditional NLP techniques and their long-standing use in computational linguistics can be traced back to the 1960s (COL, 1967). In this section, what is being considered are techniques that give no use to machine learning or deep learning methods to perform an NLP task. These techniques are often favored in social sciences due to their accessibility and ease of use (Kroon et al., 2021). In this section, traditional NLP methods are divided into two main approaches: count-based and embedding-based approaches.

Count-based approaches quantify words and n-grams to analyze the bias within the text. Spinde et al. (2020) evaluated German news articles for political bias using an IDF-based component together with a custom bias lexicon, arriving at an F1 score of 0.31 in the best case, highlighting how challenging the task of media bias detection is, and how limited traditional methods can be. De Arruda et al. (2020) applied an outlier detection framework to identify political bias in Brazilian news, an innovative idea within media bias detection literature. News would be considered biased when classified as an outlier, meaning when it deviates considerably from the average behavior. D'Alonzo and Tegmark (2022) developed a method based on Singular Value Decomposition of matrices composed of word and n-gram frequencies to position newspapers on a two-dimensional bias map. The dimensions of the bias map represented traditional left-right bias and establishment bias. Zahid et al. (2020) analyzed media bias on Twitter through a count-based model using mention counts and sentiment polarity rates. Grefenstette et al. (2004) evaluated the sentiment in news articles about public figures by identifying affective words in the text surrounding a figure's name, and calculated a sentiment score by comparing the frequency of positive versus negative words of affection. They concluded that these scores were in line with the publicly assumed bias of the analyzed news outlets. Finally, Carvalho et al. (2021) explored explainable

models for assessing news reliability to combat misinformation and improve media literacy. The study identified and evaluated credibility indicators in Portuguese news articles through a detailed annotation study, comparing mainstream and alternative media. The findings demonstrated how content features like sentiment, title accuracy, and fallacies influence credibility perception and highlighted the need for tools aiding critical evaluation of news content.

Embedding-based techniques analyze word associations through their embedding representations. Leavy (2020) detected gender bias towards political leaders in Irish newspaper coverage using TF-IDF, and other feature representation techniques such as bag-of-words and boolean-based, revealing coverage biases against female politicians. Dense embedding techniques like the Word-Embedding Association Test (WEAT) (Caliskan et al., 2017) measure token similarity to identify biases. WEAT is a statistical method developed to evaluate biases in word embeddings using a technique inspired by the Implicit Association Test used in psychology to measure implicit biases in humans. Ferrer et al. (2020) applied WEAT to analyze Reddit’s linguistic bias specifically towards gender, religion, and ethnicity. Mendelsohn et al. (2020) examined LGBTQ dehumanization in New York Times articles using embedding-based methods and found increasing humanization over a large period of time. On a similar note, Wevers (2019) studied gender bias in Dutch newspapers between 1950 and 1990 using word embeddings to analyze linguistic patterns in six newspapers with varying ideological backgrounds. To do this, it created vector representations of gender-related words and measured the distance between target words and these gender vectors. It identified shifts in gender bias over time, revealing that while some themes, like sexuality, shifted toward associations with women, others, such as money and work-related terms, gravitated toward men. Other studies, such as Papakyriakopoulos et al. (2020), using politics-related social media posts, and Kroon et al. (2021), using a dataset of 3 million Dutch news, applied embeddings to detect gender and racial biases, respectively. Spinde et al. (2021b) presented an approach for identifying biased terms in news articles by comparing word embeddings trained on text from two ideologically divergent news outlets, using the Word2Vec continuous skip-gram architecture⁷. It proposes that words associated with bias exhibit greater contextual variation across the two embedding spaces compared to neutral words. While results showed some potentially bias-inducing words among those with the largest distances, the findings were not statistically significant. In a similar way, Sales et al. (2021) used embedding-based methods to detect objectivity bias and employed the proposed method in Portuguese, Spanish, German, and English datasets.

In conclusion, traditional NLP techniques are characterized by their simplicity, serving as valid tools within Social Sciences and as useful baselines to measure performance (Spinde et al., 2023). Despite their usefulness, these techniques are usually limited in terms of their capabilities, which is something to keep in mind, especially in the context of a subjective task such as media bias detection.

⁷ Although Word2Vec for itself is machine learning-based, the overall method is here framed within traditional NLP

2.2.2 Traditional Machine Learning Methods

This section considers research on media bias detection using machine learning (ML) methods that do not involve deep learning, specifically excluding approaches based on complex neural network architectures. ML methods are usually more capable and outperform traditional NLP techniques as those from Section 2.2.1 (Spinde et al., 2023). Despite this, researchers still have to engineer features out of text since algorithms like Support Vector Machines (SVM) or Random Forest (RF) are not able to handle unstructured text data. These features are commonly extracted using computational linguistics, which can be similar to the techniques presented in the previous section. Lastly, ML methods are still widely used in today's research for their relative simplicity and explainability, meaning one can interpret the model's decisions (Spinde et al., 2023).

Looking for an automated way to detect fake news online, Pérez-Rosas et al. (2017) presented two new datasets: one focused on fake content within standard news domains such as business, politics, sports, and others, and a second dataset exclusively targeting celebrity-based news since these are more prone to rumors, hoaxes, and fake content. In terms of methodology, the features extracted from text contain n-grams, punctuation counts, and readability features, among others. The study made exclusive use of a linear SVM to classify the news, obtaining an accuracy of 0.74 on the first dataset and 0.76 on the second one, showing how the performance can be improved compared to traditional NLP. In 2018, Monteiro et al. (2018) presented one of the first datasets in Brazilian Portuguese for fake news detection, commonly referred to as Fake.Br, and trained another SVM classifier, achieving impressive performance on the task with metrics like accuracy and F1-score around 0.89. Cruz et al. (2019) used interpretable linguistics-based features to tackle the SemEval-2019 Hyperpartisan News Detection challenge (Kiesel et al., 2019). The features included document, sentence, and word lengths, use of punctuation, and n-gram-based features, among others. Regarding algorithms used, previously mentioned choices such as SVM and RF were present, with the addition of Gradient Boosted Trees (GBT), another popular tree-based algorithm. Also participating in the same challenge, Anthonio and Kloppenburg (2019) presented a linear SVM classifier using features like sentiment scores obtained from the VADER sentiment analysis tool (Hutto and Gilbert, 2014), bias-related words, and n-grams. The system relied exclusively on the intensity of negative sentiment in the text to derive political bias and achieved an F1-score of 0.694. Researching in other languages, Abonizio et al. (2020) studied news reliability in Portuguese and Spanish using available datasets, including Fake.Br from Monteiro et al. (2018) for the Portuguese case. The feature engineering phase targeted language-independent features, and on top of some of the algorithms introduced before, this study made use of k-Nearest Neighbors (k-NN) and Extreme Gradient Boosting (XGB), concluding that ensemble-based algorithms performed the best. While presenting a novel dataset for political orientation in news articles (MBIC), Spinde et al. (2021c) built an automated system for political stance detection based on features such as POS tags, syntactic dependencies, named entity types, embedding vectors, and others. The best classifier was an XGB model that obtained an F1-score of 0.43. Kameswari et al. (2020) explored the relationship between presuppositions in news articles and bias, proposing

guidelines for identifying presuppositions, and presenting an annotated dataset of 1,050 articles. The paper introduced a supervised classification method with the use of Naive Bayes (NB), Logistic Regression (LR), SVM, and RF. [Fayaz et al. \(2022\)](#) addressed the challenge of detecting fake news in the current environment of rapid growth of online content by employing several classifiers, including k-NN, RF, NB, LR, GBT, and others. The study engineered 23 linguistics-based features, selecting the best ones using techniques like Chi2 and information gain, and once again, it concluded that, generally speaking, ensemble-based algorithms perform well. Ending with European Portuguese research, [Rodrigues \(2020\)](#) put together one of the first resources for this Portuguese variety in the form of a dataset with 3,764 instances. The dataset was then used extensively with several combinations of feature extraction techniques such as Bag-of-words, TF-IDF, and word embeddings, in conjunction with popular machine learning classifiers presented before, to create an automated system for fake news detection within the European Portuguese context.

2.2.3 Deep Learning Methods

When compared to traditional ML methods, deep learning (DL) methods are less interpretable and usually more expensive. However, recent advances have demonstrated remarkable performance increases in tasks with complex domains, such as media bias detection ([Spinde et al., 2023](#); [Aribandi et al., 2021](#); [Rodrigo-Ginés et al., 2024](#)). Being less interpretable means that it is difficult to understand the model's decisions and what they are based on, and this is essentially why deep learning models are often referred to as “black boxes” ([Şahin et al., 2024](#)). In terms of performance, these neural network-based (NN) models are indeed capable of outperforming traditional ML models, but that can also bring a significant cost in terms of the amount of data needed to make a difference, and computational power for training and running the model ([Lo et al., 2024](#)).

2.2.3.1 Recurrent Neural Networks

NLP involves working with text, and text can be represented as a sequence of tokens, be it character-level tokens, word-level tokens, or something in between, like sub-word tokens. Traditional ML algorithms, and standard feed-forward NN architectures for that matter, lack capability in terms of representing sequential data, as far as keeping the order inherent to this kind of data. This is where Recurrent Neural Network (RNN) ([Rumelhart et al., 1986](#); [Williams and Zipser, 1989](#)) architectures are very useful. The RNN is very well suited for the specific task of processing sequential data, such as text, but from early on, it revealed some flaws. With longer sequences, the gradients propagate backward just up to a certain extent during the training process, which severely limits the model's capacity to capture longer-term dependencies in sequences effectively, an issue known as the vanishing gradients problem ([Hochreiter, 1998](#)). To improve performance with longer sequences, the Long Short-Term Memory (LSTM) ([Hochreiter and Schmidhuber, 1997](#)) was created as a variation of the original RNN architecture by introducing a memory mechanism that allows the network to learn patterns that occur many steps apart, with the help of three

main components: the input gate, output gate, and forget gate. These gates work together to control the flow of information into and out of the memory cell, enabling the LSTM to selectively remember or forget information as needed. The architecture of an LSTM unit is illustrated in Figure 2.4. Later, in 2014, [Chung et al. \(2014\)](#) introduced a simpler alternative to the LSTM – the Gated Recurrent Unit (GRU). Both advancements contributed significantly to improve the performance of RNN architectures when dealing with longer-term dependencies by helping to mitigate the vanishing gradients problem ([Hochreiter and Schmidhuber, 1997](#)).

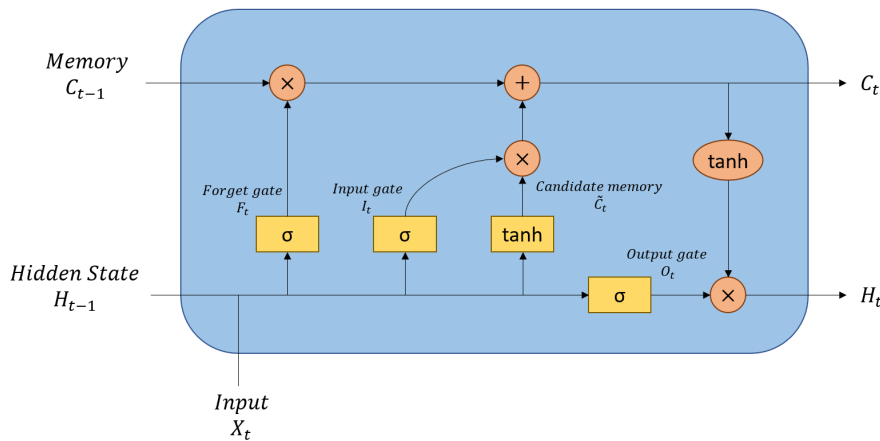


Figure 2.4: LSTM unit architecture

Due to their suitability to process text data without the need to perform specific feature engineering beforehand, RNNs and, in particular, their LSTM variant, have been widely used as a tool for media bias detection. In order to tackle the task of propaganda detection at the sentence level, [Ferreira Cruz et al. \(2019\)](#) proposed not only a GBT model with handcrafted features but also a bidirectional LSTM coupled with an attention layer⁸ in a 5-model ensemble format. The LSTM-based model outperformed the GBT model considerably, highlighting the capabilities of deep learning methods for text classification. Also in [Cruz et al. \(2020\)](#), the same authors proposed several model architectures to address the problem of hyperpartisan news detection using the SemEval-2019 dataset. The proposed models included a bidirectional LSTM, an attention-based bidirectional LSTM, and a Hierarchical attention network, the last one using a bidirectional GRU as the mechanism that processes the text sequences into sentence embeddings, and then finally into document embeddings in a two-step process. The best-performing model resulted from an ensemble of hierarchical attention networks and achieved an accuracy of 0.825. [Hong et al. \(2023\)](#) introduced a novel approach to detect political bias in news articles by disentangling writing style from structural elements to improve generalizability and reduce overfitting to particular writing styles of news outlets. It employs a multi-head hierarchical attention model with a bidirectional LSTM that captures sentence-level semantics and document-level rhetorical structures to classify bias. Once more, the model demonstrates superior performance compared to traditional methods,

⁸The concept of attention will be introduced in Section 2.2.3.2

especially in distinguishing bias across diverse news outlets. Additionally, the model identifies key sentences in articles, enabling better interpretability. Contributing with another dataset to classify news articles as left-wing, center, or right-wing biased ideologically, [Baly et al. \(2020\)](#) also trained an LSTM-based model for the task. On top of the dataset, another contribution was the evaluation of techniques for eliminating the effects of media outlet language characteristics from the training process. The research proposed adversarial media adaptation and triplet loss to remove linguistic characteristics from the training data. The techniques for improving training effectiveness improved the model's classification results.

2.2.3.2 Attention and Transformers

The same year GRU was introduced, [Bahdanau et al. \(2014\)](#) presented the attention mechanism. This was an important breakthrough in the field of NLP because it allowed the model to learn something really interesting. To give an example, when using an encoder-decoder architecture for Machine Translation, another popular NLP task, the encoder usually outputs an internal state in the form of a single vector that is based on the representations of each token of the input sequence. The decoder uses this encoding vector to output the intended sequence in the translated language. What attention brings to this process is the ability for the decoder to attend, that is, give higher weight to specific encoder representations of individual input tokens, which means that each decoding step has a different representation of the encoded sequence and allows the model to learn more complex relationships between words.

In 2017, another important breakthrough emerged in the form of the Transformer architecture ([Vaswani et al., 2017](#)). The original architecture is presented in an encoder-decoder format and employs a variation of the original attention (scaled dot-product attention) as its main mechanism. Other tricks include positional encoding, where input order is kept within the token representation itself, and teacher forcing, which acts on the decoder side during training by inputting the ground truth token into the decoder (which inherently works autoregressively) instead of the decoder's output from the previous step. All this together allows for the training of a Transformer model to work in a fully parallelized way, completely dropping the recurrence (sequential) logic that is native to RNN architectures, speeding up the training process in a dramatic way.

Based on the transformer architecture, many adaptations and variations have emerged, tweaking the original architecture into several pre-trained models. Examples of this are XLM ([Lample and Conneau, 2019](#)), an encoder-based transformer, T5 ([Raffel et al., 2019](#)) and BART ([Lewis et al., 2019](#)), both encoder-decoder transformers, or even GPT-3 ([Brown et al., 2020](#)) and LLaMA ([Touvron et al., 2023](#)), decoder-based transformers. There is a particular adaptation that deserves a more detailed look due to outperforming these other encoder-decoder or decoder-only models, particularly in media bias detection tasks ([Spinde et al., 2023](#)), and also for the extensive research usage: Bidirectional Encoder Representations from Transformers, or for short, BERT ([Devlin et al., 2018](#)). As the name indicates, it is a language representation model based on the encoder side of a transformer that performs exceedingly well on several NLP tasks. BERT is trained in two different ways in a self-supervised environment: 1) it employs a masked training method, where

the input is text with 15% of the words missing and the task is to predict the masked words; and 2) given two different sentences, the model aims to predict if one follows the other (next sentence prediction). Using its bidirectional architecture, these techniques allow it to produce very rich context-based representations of the words, as well as a condensed sequence representation in the form of the output relative to the [CLS] token, which is very useful for document classification tasks. Figure 2.5 illustrates the use case where BERT can be fine-tuned for sentence classification, as presented in Devlin et al. (2018).

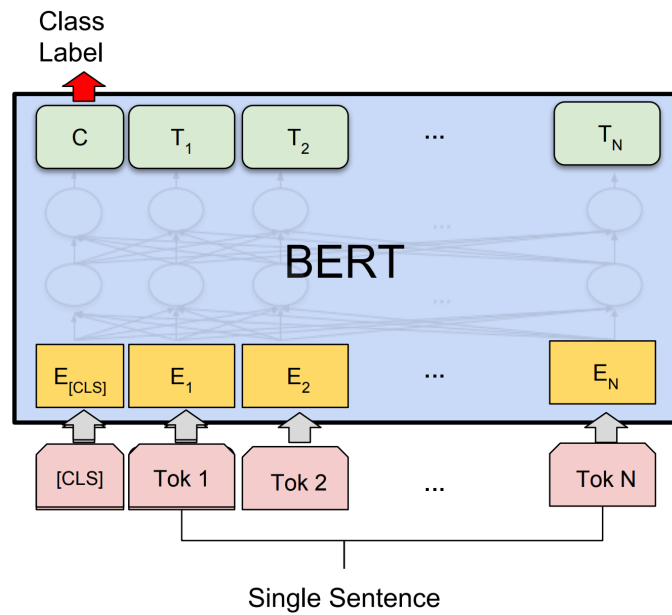


Figure 2.5: BERT for text classification (from Devlin et al. (2018))

BERT-based models have become a prominent way of performing several NLP tasks, including text classification, due to their effectiveness and improved performance compared to RNN-based models (Spinde et al., 2023; Rodrigo-Ginés et al., 2024). In an interesting twist to the majority of research that has been cited until now and more aligned with the bias correction (Hamborg, 2022) paradigm from Section 2.1.3, Pryzant et al. (2019) developed methods for automatically neutralizing subjective bias in text, transforming biased language into neutral to enhance objectivity in media or any other texts. The authors introduced a dataset of 180,000 bias-neutral sentence pairs sourced from Wikipedia edits. They proposed two sequence-to-sequence models: a MODULAR model that detects and edits subjective words and a CONCURRENT model that integrates bias detection into text generation, both using BERT-based architectures. Large-scale human evaluations demonstrated that these models effectively reduced bias across various domains, such as news and political speeches, outperforming traditional style transfer and machine translation techniques. Dadu et al. (2020) focused on detecting subjectivity bias at the sentence level by using contextualized word embeddings, with applications in tasks like sentiment analysis and propaganda detection. In terms of methodology, the study explored several BERT-based models, such

as RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2019), including also ensemble techniques, optimizing their performance for detecting bias at the sentence level. The ensemble model significantly outperformed baseline models like BERT-large and a bidirectional LSTM, achieving an F1-score of 0.7, highlighting the robustness of ensemble approaches in a subjective task such as bias detection. With a similar objective of detecting text subjectivity, Huo and Iwaihara (2020) applied fine-tuned BERT-based classifiers and BERT in conjunction with LSTM networks in several datasets, finding that a fine-tuned BERT classifier performed the best. Another BERT-based model proposed by Kim and Johnson (2022) is CLoSE. The model uses contrastive learning to predict political bias in news media by identifying frames and subframes, focusing on politically polarized issues like abortion, gun control, and immigration. It was trained on a custom triplet dataset constructed from news articles with annotated subframes and bias labels. Regarding performance, CLoSE outperformed baseline BERT models in both subframe detection and political bias classification. Raza et al. (2024) introduced Dbias, a Python library designed to detect and mitigate biases in textual data, particularly in news articles, using a four-phase pipeline. This pipeline consists of detection (with DistilBERT (Sanh et al., 2019)), recognition (utilizing RoBERTa), bias masking, and debiasing, and is fine-tuned on the MBIC dataset (Spinde et al., 2021a). Dbias identifies biased words, masks them, and suggests alternative sentences with bias-free or less biased terms, maintaining semantic integrity. Introducing a novel multi-dimensional framework and a new dataset for analyzing political ideology and polarization in news articles, Sinno et al. (2021) focused on the ideological content of discussed policies rather than the writer's stance. The dataset is annotated at the paragraph level, spanning three dimensions – economic, social, and foreign policy – by trained political scientists to allow for a comprehensive analysis of ideological shifts over time. The study also evaluates several computational approaches for ideology prediction, finding that fine-tuned BERT models performed best, outperforming other deep learning methods like a 2-layer bidirectional LSTM and even other transformer-based systems, corroborating other authors' claims already cited. In an interesting application of BERT-based models, Zhou et al. (2021) studied the construction and use of datasets for unreliable news detection, emphasizing their potential impact while revealing critical limitations that may degrade generalization to realistic settings. The research employed RoBERTa for this task, demonstrating that many models perform well, not by genuinely addressing unreliable news detection, but by exploiting other factors like sources and selection bias. They showed that in a clean split with no train/test source overlap, model performance dropped by over 10%, highlighting the need for improved dataset creation methods to reflect real-world complexity. It is a critical limitation to keep in mind when developing any tool or system as the one this dissertation focuses.

For research working within the European Portuguese landscape, there are three studies worth mentioning. Moura et al. (2021) proposed a system for automatic detection of fake news and started by assembling a dataset from Portuguese media, labeling each article as fake content or not depending on the source that produced said article, arriving at a total of 110,066 credible news articles and 10,343 fake news articles. On top of performing feature extraction in the form of n-grams, frequencies, metadata, and readability metrics to feed into traditional machine learning

algorithms, the study also proposed a BERT-based model that achieved impressive performance with an accuracy of 0.97. Another very complete research piece is that from [Tavares \(2023\)](#). It was already introduced in Section 2.1.4.4 when referring to a prototype visualization tool similar to Ad Fontes Media’s offer. In terms of development, the research used the output of BERT-based models such as Albertina ([Rodrigues et al., 2023](#)), BERTimbau ([Souza et al., 2020](#)), and XLM-RoBERTa ([Conneau et al., 2019](#)), in conjunction with other features extracted from the news articles to classify them according to their reliability, objectivity, and political bias. The problem is essentially multi-class, but because of the lack of data covering the whole scope, it was tackled by treating each framing axis as a standalone problem. For reliability, the dataset was the one put together earlier by [Moura et al. \(2021\)](#) for fake news detection in Portuguese; for objectivity, BABE ([Spinde et al., 2022b](#)), a popular English dataset labeled for political bias and objectivity was translated into Portuguese using DeepL; and for the political bias detection part the research used a proxy dataset by scraping Portuguese parliamentary minutes, which is not ideal because of the hardly comparable domain of news articles and political debate. Lastly, [Afonso and Rosas \(2024\)](#) constructed a complete and varied dataset labeled for reliability classification that was employed in extensive experiments. From traditional machine learning methods such as XGB, RF, GBT, SVM NB, LR, and more, to deep learning methods like BERT and other variations, unidirectional LSTM, and bidirectional LSTM, the research experimented with a wide variety of possible training methods with different sets of features. Interestingly, in this case, traditional ML techniques were capable of keeping up in performance with their DL counterparts.

In conclusion, DL methods stepped up performance on tasks that deal with intricate and complex data distributions, such as text and the task of text classification. Even within DL methods, there is a variety of possibilities that range from older, yet effective, architectures such as the LSTM, to recent advancements that seem to have overtaken most NLP applications, such as transformers and, in particular, BERT-based models.

2.2.4 Prompting Methods and Optimization

After the emergence of large language models, particularly in their now common decoder-based transformer form, a new paradigm within NLP research has emerged. In 2020, [Brown et al. \(2020\)](#) introduced GPT3, and showed the wide variety of tasks that a language model exclusively trained on the task of next-word prediction could solve. This concept essentially introduced a new research field commonly referred to as prompt-based learning ([Liu et al., 2023](#)). Whereas supervised learning methods, like the ones presented in previous sections, tune a model to map an input x and predict output y , prompt-based learning is based on the probability of text as estimated by large language models.

In its simplest version, a prompt-based technique takes the original input and inserts it into a prompt template, leaving empty space and allowing the language model to fill it, performing the task it was trained for. From the model’s prediction of the remaining text, output y can be extracted. To give a simple sentiment analysis example, given the input sentence “This dissertation is very good!”, a prompt template could be “This dissertation is very good! I felt __”, leaving the model to

complete the sentence by filling the last slot (Petroni et al., 2019; Schick and Schütze, 2020). After this, the output of the model can be mapped into one of two classes, for instance, “positive” or “negative”. It is worth noting the flexibility of this simple prompt template in the form of “[X] I felt ___”. Like this, it allows pretty much any input sentence x to be classified in terms of sentiment, just as with any supervised learning method. This brings several advantages versus fine-tuning pre-trained models in a supervised context, given that large-scale language models were shown to be able to perform well on various NLP tasks (Brown et al., 2020; Sun et al., 2021; Ouyang et al., 2022), while also reducing the specific need of large labeled datasets (Liu et al., 2023).

The simple example given above can be described as a discrete prompt (Liu et al., 2023). This is the simplest kind of prompt and has the convenient property of being human-interpretable. Choosing the best discrete prompt can be a challenge in itself, and there is interesting research in this field. Jiang et al. (2019) automatically found templates given a set of labeled training examples. The method was based on scraping a large text corpus of text, e.g., Wikipedia, for strings containing the text and respective label, and finding either the middle words or dependency paths between the inputs and outputs. With frequent middle words, for example, the prompt template could be “[X] [middle words] ___”. An interesting research by Guo et al. (2023) presented a novel framework that integrates LLMs with Evolutionary Algorithms (EAs) for automatic discrete prompt optimization as a derivative-free approach to address this issue. The framework uses evolutionary operations like mutation and crossover, with LLMs acting as evolutionary operators to maintain coherence and readability in prompts. Across a wide range of tasks and datasets, it outperforms human-designed prompts and prior automated methods, achieving up to 25% improvement. Other optimization techniques rely on gradients, for example Wallace et al. (2019) applied a gradient-based search over actual tokens to find prompts that can trigger the model to generate the desired target prediction. Another popular method for discrete prompt optimization is to have a baseline prompt created beforehand, and then create a number of different prompts that essentially say the same thing with different words – paraphrasing. The original and derived prompts can then be tested using a validation set (Jiang et al., 2019; Haviv et al., 2021).

A prompt is nothing else than an input to an LLM, and understandably, it highly influences the performance of the model when performing the desired task (Liu et al., 2023; Zhu et al., 2023). The prompt can be conveniently interpretable, like a discrete prompt, but it can also be uninterpretable and continuous. Continuous prompts are also sometimes referred to as soft prompts. This removes the constraint for the embeddings of the prompt template to be embeddings corresponding to natural language words (Liu et al., 2023). In this case, the prompt template has its own parameters that are trainable with the use of labeled data and backpropagation, so as to optimize the LLM input according to the task. Li and Liang (2021) called it “prefix tuning” and presented a method where a sequence of vectors is added to the input, and then supervised training is employed in order to optimize just these added vectors, leaving the LLM parameters frozen. The authors highlight the high sensitivity of continuous prompts to initialization. Following up, Zhong et al. (2021) experimented with initializing the prompt template in a discrete fashion, almost serving as a heuristic in the search for an optimal prompt, and then fine-tuning its embeddings. The

research found that this approach works well and provides a good initialization point.

As noted previously, prompt-based methods can be used by simply prompting a pre-trained LLM without any specific additional training on the task, but training can increase the model's performance significantly. Classifying different methodologies according to the kind of training they employ, [Liu et al. \(2023\)](#) proposed four training strategies when using prompt-based methods:

1. **Tuning-free prompting:** Also known as in-context learning ([Dong et al., 2022](#)), this is the simplest iteration. It is based on keeping the LLM's parameters frozen and just prompting the model to perform the desired task. It can be done in a zero-shot fashion, as in the example given above, or using what the authors call "prompt augmentation" - or few-shot learning - a technique in which some examples are provided in the prompt to induce the model to an even greater extent into what the output should be. Although this iteration is simple and cheaper to execute, it might require a lot more work in terms of prompt engineering to get the model to execute the task with high accuracy.
2. **Fixed-LM Prompt tuning:** Prompt tuning techniques have also been explored in some detail for continuous prompts and represent cases in which additional parameters are added at the prompt level. Although the LLM parameters are kept frozen, the prompt-related ones are trained in a supervised context. This has the advantage of retaining the whole knowledge from the LLM intact, while also increasing the model's performance on the task.
3. **Fixed-prompt LM tuning:** This third iteration is the most similar to methods presented in Section 2.2.3.2 as the parameters to be tuned are from the model itself, either in their totality or just a portion of the trainable parameters. This allows for efficient learning on the exact task but can hinder the model's general capabilities on other tasks and also has the obvious disadvantage of high computational cost.
4. **Prompt + LM fine-tuning:** Finally, the last strategy not only adds trainable prompt parameters but also fine-tunes the model's parameters, essentially merging the previous two strategies. It could bring the best results, but also at the greatest cost.

2.2.5 In-Context Learning

After introducing a variety of prompt-based methods in the previous section, this section focuses on a specific one while also reviewing some research on practical applications. In-context learning (ICL) is an emerging paradigm in NLP that enables large language models to make predictions based solely on a few examples provided as part of the input context, without requiring explicit parameter updates ([Dong et al., 2022](#)). This approach, initially explored by [Brown et al. \(2020\)](#) with the introduction of the GPT-3 model, represents a significant shift from traditional supervised learning paradigms as the ones presented in previous sections. In ICL, the model uses natural language templates containing examples as demonstrations to infer patterns and solve tasks through analogy. The core idea of ICL involves assembling a prompt that contextualizes the model into the task with a small number of labeled examples, or even in a zero-shot fashion. Unlike fine-tuning

methods that adjust model weights, ICL operates in a context that does not require any parameter adjustments, leveraging the model's pre-trained knowledge of language to solve several other end tasks. This capability has been shown to excel in tasks ranging from mathematical reasoning to compositional generalization (Wei et al., 2022a,b).

Zero-shot learning, a form of ICL, is the ability of a model to perform tasks it has not been explicitly trained on, just by using general knowledge acquired during its development stages and task instructions. Qin et al. (2023) studied ChatGPT's zero-shot learning capabilities across a diverse range of NLP tasks. The methodology involved evaluating ChatGPT's outputs using standard metrics like accuracy and ROUGE scores across 20 datasets, with experiments featuring chain-of-thought prompting for reasoning tasks. The research concluded that despite its promise as a generalist model, ChatGPT seems to struggle overall in specific domains and lags behind fine-tuned models. Continuing with general-purpose research on ICL, Cahyawijaya et al. (2024) explored few-shot ICL capabilities for low-resource languages. A key finding was that few-shot approaches improve performance significantly over zero-shot setups, but highlighted the importance of task-specific quality examples, rather than random selection. Out of the domain of media bias detection, but providing very relevant insights in the context of this dissertation, Carpenter et al. (2024) explored the use of large language models for assessing student self-explanations in the context of Computer Science tests, in an attempt to improve the efficiency of manual assessment in large classrooms. To address this task, the study compared prompting techniques on models like GPT-3.5, GPT-4, and LLaMA 2 against fine-tuned FLAN-T5. As far as prompting is concerned, the authors experimented with zero-shot approaches, as well as few-shot in 4 variations: adding rubric items for the question, adding an exemplar response, adding other students' labeled responses, and everything together. Results suggested that few-shot approaches generally improve performance, but overly detailed exemplar responses may not be the best option.

Considering studies that employ ICL for tasks analogous to those addressed in this dissertation, Golan et al. (2023) evaluated ChatGPT's ability to assess the quality and readability of online medical information, using established tools as benchmarks. The study found significant discrepancies between ChatGPT's assessments and those from human reviewers or established tools, with ChatGPT often overestimating quality and underestimating readability levels. Despite its efficiency and consistency, the findings suggest that ChatGPT's current capabilities are insufficient for reliable evaluation. Naous et al. (2023) introduced README++, a multilingual and multi-domain dataset for readability assessment in five languages, namely Arabic, English, French, Hindi, and Russian. It includes 9,757 human-annotated sentences. The paper benchmarks large language models for readability assessment using few-shot prompting and supervised methods. Few-shot experiments employed models like GPT-3.5, GPT-4, and Llama variants, where the models were provided with five in-context examples sampled from the dataset to assess readability. Results demonstrated that increasing domain diversity in few-shot examples significantly improved performance, achieving correlation values over 0.7 in the best cases. Few-shot approaches, however, underperformed compared to fine-tuned models such as BERT, T5, and XLM, suggesting there is room for improvement with prompting techniques. Readability assessment and text summarization seem to be

particularly popular tasks in scientific research when employing pre-trained large language models. On this, [Rodrigues \(2024\)](#) used article lead section pairs from English Wikipedia and Simple English Wikipedia for 9 different categories to assess the capability of LLaMA 3 8B to accomplish the two tasks. The readability assessment task was framed as a classification problem, where the model is prompted to rate the lead section on a discrete grade scale in the range of 1-20. The evaluation was based on comparing the model's performance against standard linguistics metrics for readability assessment, as well as the percentage of cases where the Simple English Wikipedia version was ranked at the same or lower rating as the English Wikipedia counterpart. Regarding results, 83% of the model's assessments deviated by under three grade levels from standard metrics, and it consistently rated Simple English Wikipedia sections as simpler than their standard counterparts.

Addressing media bias more directly, [Alghamdi et al. \(2024\)](#) proposed a prompt-based learning framework for fake news detection using GPT-2 ([Radford et al., 2019](#)) as the foundational language model, focusing on zero-shot and few-shot learning. Regarding methodology, the authors employed prompt engineering, where templates and verbalizers are designed to guide the model's predictions. Additionally, domain information was incorporated into prompts to provide context. Sensitivity analyses revealed that slight changes in prompt design and wording can affect performance, highlighting the importance of prompt optimization. An important insight was also that incorporating minimal labeled data improves accuracy, showcasing the advantage of few-shot approaches versus zero-shot. The results suggested that the proposed framework is effective for fake news detection, offering a scalable solution for scenarios with limited labeled data. Another study researching reliability detection in news content with the use of prompt-based techniques is [Wen and Younes \(2024\)](#). This paper evaluated ChatGPT's zero-shot capability to detect six types of bias-related characteristics, comparing its performance to fine-tuned models like BART, ConvBERT ([Jiang et al., 2020](#)), and GPT-2. In terms of prompt engineering, the prompts themselves were generated by ChatGPT on several iterations and then tested on a validation set to choose the best-performing prompt. In terms of solving the task in a zero-shot manner, ChatGPT performs respectably on racial or context biases but underperforms fine-tuned models in fake news detection, probably due to the complexity and ambiguity of the task, according to the authors. The paper concludes by highlighting the potential for scalability when using zero-shot approaches, but underlines limitations in detecting nuanced biases without specific training. Another popular bias-related task is political stance detection, as already shown in previous sections. [Lin et al. \(2024\)](#) presented IndiVec, a very interesting framework for political bias detection by adapting ICL methods. The methodology essentially relies on a vector database and bias indicators generated by the model across dimensions like tone, language, and citations. When analyzing a new text, IndiVec matches descriptors derived from the input text to the stored indicators, assigning bias labels based on a majority voting system. It excelled in adaptability to different domains and improved explainability and performance versus simpler prompt-based methods. In Section 3.3.2, this framework is explored in greater detail.

2.3 Research Gaps

Throughout this chapter, an extensive literature review has been presented covering not only general media bias research, but also several text classification techniques and respective applications targeting media bias detection. From this review, there are a few important takeaways to help determine how this dissertation can meaningfully contribute to the existing research.

As it was shown, a lot of work has been done covering different media bias types and similar concepts. Despite this, the vast majority of research targets the English language, understandably, and particularly American news, meaning that any other Western language pales in comparison in terms of targeted research. European Portuguese, in particular, has very few resources in terms of available datasets and is one of these poorly covered languages. Covering the Portuguese media landscape is one of the ways in which a good contribution is delivered. Additionally, most of the research in Portuguese covers media bias from a reliability / fake news detection perspective, leaving uncovered many other bias-related topics. The only research found covering, for instance, political stance in the Portuguese media (Tavares, 2023) does so by using parliament interventions from Portuguese political parties as a proxy dataset, since there are no labeled datasets for this task. Given that the data is just a proxy for politically biased text, there is certainly room for improvement on this front that this dissertation explores. Political bias is a particularly interesting topic to research in multiple geographies because of cultural differences. For example, a right-wing biased piece of news can look very different in Portugal and in the United States due to inherent differences in each country's political systems.

Another component has to do with the methods employed. As previously stated, prompt-based methods are relatively recent, with the first contributions tracing back to just a few years ago (Brown et al., 2020). Although recent, these methods have proven useful, including in media bias detection contexts. Once again, the application of prompt-based methods in the Portuguese context is almost non-existent, which reveals another gap. Furthermore, existing research usually makes use of extremely large models such as GPT4 when it comes to prompt-based methods. During the development of this dissertation, the aim was to use smaller models and assess how capable they are at the task, which ultimately would allow the tool to be run locally with relative ease.

Lastly, the development of an interactive visualization tool to represent media articles in a bias-representing space is also something worth considering as a relevant contribution, for its exclusivity and current limitations of existing similar solutions in terms of axis variety. Commercial-grade tools such as those offered by AllSides and Biasly only classify news in a 1-dimensional space. Ad Fontes Media's tool provides an additional dimension, but in the case of this project, the target was to develop a set of 4 axes from which the user can choose, adding some degree of variety and customization.

Chapter 3

Research Methodology

This chapter presents every detail about the methodology employed during the development of this dissertation. Firstly, details about the axes chosen to frame each article are provided. Furthermore, some of the existing datasets are presented, as well as the choices made on this front and the reasons behind them. Lastly, the actual methodology for news classification is detailed thoroughly, including the classification strategy, framing, and prompt optimization, backed by the literature already presented.

3.1 Framing Axes

Most existing research focuses on one specific aspect of media bias, be it the detection of fake news, the detection of political stance, or other bias-related concepts (Spinde et al., 2023). This dissertation, however, looks into several types of bias-related topics, offering a variety of perspectives from which news articles can be assessed. The current section presents each concept that is explored in detail over the course of this research.

Political Bias Classifying news according to their political leaning is an interesting task, given that it depends on the context in which it is performed. Conclusions from studies investigating political bias in the American context, for example, do not necessarily apply to the Portuguese context. The media outlets are different, many topics of discussion are different, and most importantly, the overall ideology is different on a country-wide scale. Through this research, Portuguese news consumers gain insight into how favorable different media outlets operating in Portugal are towards specific sides of the local political spectrum.

Reliability Assessing the trustworthiness of news is possibly the most obvious route to take. It is one of, if not the most researched factor when it comes to media bias classification, and, possibly for that reason, it is also the one with the largest amount of resources available. To be clear, this

reliability factor is assessed exclusively through the linguistic characteristics of each article, based on labeled datasets suitable for the task, and not by fact-checking what the article states against any external knowledge base.

Objectivity As mentioned in Section 1.1, journalism should follow several criteria in order to add value to society. One of these criteria is to adopt objective reporting practices, meaning a lack of opinionated content when reporting actual events. The presence of opinion-based factors within the context of a supposedly objective news piece is measured in this Objectivity axis.

Readability The fourth and last framing component is readability. This is perhaps the least bias-like characteristic when compared to the rest, but is also something interesting to be assessed in the context of news text. Additionally, it also brings some value to observe the relationship between this and the other axes.

3.2 Datasets

Throughout Chapter 2, several research pieces were presented that, in addition to interesting applications and methodologies, also introduced new datasets, contributing to a specific branch of media bias research. There are many available datasets covering the English language and targeting different types of bias. Examples of this are BABE (Spinde et al., 2022b), MBIC (Spinde et al., 2021a), the NELA-GT series of datasets (Gruppi et al., 2022), FlipBias (Chen et al., 2018), MBC (Card et al., 2015), BASIL (Fan et al., 2019), ABP (Baly et al., 2020), SemEval-2019 (Kiesel et al., 2019), MBIB (Wessel et al., 2023), ReadMe++ (Naous et al., 2023), and lastly FakeNewsAMT and Celebrity from Pérez-Rosas et al. (2017). All of these examples together cover issues like fake news, subjective and opinionated news content, politically biased news articles, and more.

In Portuguese, the field is not nearly as rich in resources, and since there is no dataset capable of covering the whole scope that this dissertation covers, the only solution is to use different datasets to develop the system for the different bias-related axes, very much the same as Tavares (2023). There is Fake.Br (Monteiro et al., 2018) covering fake news content, but in Brazilian Portuguese, which is not ideal for what this dissertation targets. As previously mentioned, there are also the datasets created by Moura et al. (2021), Rodrigues (2020), and Teixeira (2021), all suited for fake news detection and in European Portuguese, but none are publicly available at the time of writing. Fortunately, two publicly available datasets serve to tackle media bias classification for two of the axes presented in Section 3.1.

3.2.1 Political Bias Dataset

Starting from the axis for which there is no available dataset, political bias proved to be the most challenging to begin with. Where one could argue that a translated dataset could be a solution for other axes, such as objectivity (Tavares, 2023), in this case, that could not be argued due to the difference in the political discourse of different geographies. There is only one research study

(Tavares, 2023) that attempted to tackle the same task for the Portuguese landscape. To solve the issue of lack of data, a proxy dataset in the form of interventions in the Portuguese parliament was used. Since political interventions and news reporting seem to belong to very different domains, a new solution is presented here.

A new dataset was created for this task based on opinion articles specifically written by personalities with some kind of direct relationship with a political party. With this being the case, an article written by someone who belongs to a communist party, for instance, can be assumed to contain slants and concepts that are associated with left-leaning ideology. The political leaning of the article is inferred based on authorship (similar to the “by publisher” dataset annotation in Kiesel et al. (2019)). Although this is not a perfect solution, it seems to be closer to the targeted domain of news articles: the publishing platforms are the same, the type of language is not much different, and most importantly, many times the topics of discussion are similar. Additionally, with one of the chosen methodologies for actual article classification, these slight differences are further mitigated to a point where it might actually not make much difference to have labeled news articles or labeled opinion articles. This is further explained in Section 3.3.2.

The data collection was performed using Beautiful Soup, a popular web-scraping library available for Python. The main challenge was finding authors with a factual political affiliation, and this was done in a manual fashion by visiting popular newspaper websites and scanning through the opinion articles section. Table A.1 in Appendix A details the list of 22 authors, their political leaning, as well as the number of articles contributed by each. The end result was a dataset with 6,901 opinion articles: 3,914 right-wing biased and 2,987 left-wing biased. Table 3.1 shows the number of articles by source and class.

Table 3.1: Political bias dataset source and class distribution

Sources	Left	Right
Expresso	2,442	3,797
Público	545	19
Eco Sapo	0	69
Dinheiro Vivo	0	29

3.2.2 Reliability Dataset

Afonso and Rosas (2024) constructed a dataset exclusively targeting the problem of fake news detection within the European Portuguese context, and since it is publicly available, it is used in the development of the reliability axis. Although no formal name is given to this dataset, it will be referred to as Fake News PT from now on for convenience purposes. It is a fairly large and balanced dataset comprised of 31,716 credible news articles and 31,520 fake news articles, with both sides sourced from multiple news distributors and websites, adding variety to the domain. At its initial state, the available version of this dataset is very noisy, with some duplicate entries

and a large number of unique sources – 3,102 to be exact – with many of them having just one entry. Furthermore, some nonsensical examples are coming from websites such as *Polígrafo*, the Portuguese fact-checking website already introduced in Section 2.1.4.4 that reports on news and publications, but does not produce news itself. Due to this, the dataset underwent a cleaning process in the form of source selection for the ones that had a meaningful representation, and also where the text corpus was composed of actual news. After removing duplicate entries, a random sampling process was done, and the final version of the dataset ended up with 10,000 news articles, half represented by reliable news and the other half by fake news. Table 3.2 shows the distribution of the whole dataset by source.

Table 3.2: Fake News PT source distribution

Sources	Reliable	Fake
SIC Notícias	2,558	0
Diário de Notícias	951	0
Expresso	787	0
Jornal de Notícias	531	0
SOL	173	0
Direita Política	0	1,833
Magazine Lusa	0	1,075
Tuga Press	0	1,023
Bombeiros24	0	844
Semanário Extra	0	225

3.2.3 Objectivity Dataset

Mainstream and Independent News Text corpus, or MINT (Caled et al., 2022) for short, is a complete dataset that groups news articles into five classes: hard news, soft news, opinion, satire, and conspiracy. The authors’ objective was to create a dataset that would be suitable for a more nuanced look at what fake news classification usually entails. Arguing that true/fake news is too strict in terms of labeling, the authors provide a more granular look, distinguishing conspiracy from satire, and what should be considered as hard, factual news, from soft, gossip-related news. In terms of basic statistics on the available data, MINT’s classes and respective distribution are shown in Table 3.3.

In terms of the use case for the dissertation, MINT is not considered for the reliability axis. The most extreme version of what could be seen as a subjective news piece is an opinion-based article and, with this in mind, the data used to build the system and evaluate its capability to frame an article into the objectivity axis is composed of the hard news and opinion examples from MINT, both representing each opposite of the one-dimensional spectrum between what is objective news

Table 3.3: MINT’s class distribution

Class	Number of Instances
Hard News	6,000
Soft News	6,000
Opinion	6,000
Satire	1,029
Conspiracy	1,249

and what is subjective news, respectively. Table 3.4 shows the distribution of sources from the two subsets, highlighting the balanced and varied composition of the dataset.

Table 3.4: Hard news and opinion articles source distribution

Sources	Hard News	Opinion
Jornal de Notícias	667	859
Expresso	667	722
Correio da Manhã	667	193
Público	667	859
Diário de Notícias	667	858
Lusa	667	0
Jornal de Negócios	666	858
TSF	666	243
SIC Notícias	666	507
Visão	0	859
Notícias Viriato	0	42

3.3 News Articles Classification

Each of the four axes proposed to frame news articles is an independent problem. Whereas for reliability, objectivity, and political bias, the problem is laid out as a binary classification problem, for readability, however, it is framed to be solved with the use of pre-established metrics and no language modeling involved. As far as the binary classification for the first three is concerned, two methodologies were followed: zero-shot prompting and a vector database approach. Both methods rely on interaction with a large language model, and in this dissertation, they are evaluated using two different models. EuroLLM 9B (Martins et al., 2024) is the result of a project co-funded by the European Union to create a series of language models that support the 24 official European languages and a few additional ones. It is open-weight and fully available in 1.7B and

9B versions; a preview version of a 22B model was released during June 2025. Since the model was specifically trained to support European Portuguese, it seems like the ideal solution for this dissertation. The second model is LLaMA 3.3 70B (Touvron et al., 2023), a medium-sized model offered by Meta AI, which is one of the most popular open-weight solutions. The idea is to compare the substantially smaller EuroLLM to Meta’s presumably more capable model.

From the methods employed, only the vector database approach requires a training set. For the three axes leveraging this methodology, only 80% of the corresponding dataset is used to build each vector database, while the remaining 20% is used to test. In the case of the zero-shot methodology, only the test sets are considered. The splits were performed randomly, keeping class distribution in both sets, and are always the same regardless of the experiment at hand for a fair comparison.

This section is divided into three parts, detailing the two methodologies for binary classification, as well as the readability assessment approach, therefore covering all four axes.

3.3.1 Zero-shot Classification

In the context of large language models, zero-shot learning, as already introduced in Section 2.2.5, refers to the ability of a model to perform tasks it was not specifically trained for. Instead, the model relies on its general understanding of language and concepts acquired from extensive pre-training on diverse data. The model circumvents the need for specifically labeled data by leveraging semantic representations, natural language instructions, and task descriptions. This paradigm has been shown to generalize well in a wide range of downstream tasks (Brown et al., 2020).

Using this method, the model is prompted directly with instructions corresponding to the desired classification criteria. In large language model APIs, two distinct types of prompts can be configured: the system prompt and the user prompt. The system prompt establishes the role, behavior, or overarching goals of the model during the session, effectively shaping its overall response style and decision criteria. The user prompt, by contrast, is the direct instruction or query given to the model for a specific instance of a task.

As the focus of this research is on Portuguese media, all articles analyzed are written in European Portuguese, and accordingly, the prompts used for zero-shot classification are also formulated in this same language. In the objectivity classification task, for example, the system prompt is used to set the model’s high-level orientation toward the task, emphasizing consistent formatting and instructing it to focus solely on formal linguistic and structural features. The translated version of the system prompt employed in this specific task is as follows:

You are an assistant specialized in assessing the objectivity of articles based on their language, structure, and transparency. Your task is to classify the objectivity and present the response in a well-defined JSON format.

Guidelines:

*- Consider **only formal elements** of the text (language, terminology, tone, structure).*

Do not assess the truthfulness of the content.

- The answer should have the following format: {"class": "given classification"}
 When you receive an article, analyze it and assess its objectivity level, carefully following the specified format.

Complementing this, the user prompt provides detailed task-specific instructions and definitions to guide the model’s decision-making for each individual article. The following is an English translation of one such user prompt used in the objectivity classification task:

*You are a model trained to classify the objectivity of opinion and news articles. Your task is to classify the provided article as "objective" or "subjective", based on its language, structure, and transparency, without assessing the veracity of its contents. An objective text tends to use neutral language, avoids expressing opinions, presents verifiable data, and maintains an impartial tone and clear structure. A subjective text tends to use opinionated or emotionally-charged language, expresses value judgments, favors a point of view without presenting counterpoints, or makes assumptions not supported by verifiable data. The output should always be a JSON with the following format: {"class": "objective"} or {"class": "subjective"} The article below is **clearly classified** with one of those options. Classify the objectivity of the following text and output **the answer in the correct format only**:*
 {ARTICLE BODY}

The design of the prompt plays a crucial role in guiding the model’s output in zero-shot settings (Alghamdi et al., 2024). Each prompt used in this study includes three essential components: contextual information outlining the classification task, clear definitions or indicative features of the target categories (e.g., characteristics of objective versus subjective language), and explicit instructions on the desired format of the model’s response. In particular, the output is requested in a JSON-style structure, which provides a straightforward and effective method for extracting and processing the model’s responses in a structured and programmatically accessible manner. This format facilitates downstream integration with the rest of the evaluation pipeline and supports consistency across multiple classification tasks.

Very similar prompts were developed for the political bias and reliability tasks. These prompts follow the same general structure – combining task context, classification guidelines, and output formatting instructions – although their specific content is tailored to the unique characteristics of each classification dimension. For brevity, the detailed examples of these additional prompts are not included in this section, but are available in Appendix B.1.

To sum up, this methodology allows for flexible and scalable classification without the requirement of annotated training data to be used. The only labeled data needed is purely for evaluation purposes – the test sets – which is particularly advantageous when working with resource-scarce languages or domains, which is the case for European Portuguese.

3.3.2 Vector Databases

Vector database techniques encompass a set of methods designed to efficiently store, index, and retrieve high-dimensional vector representations of data, in this case, text data. These embeddings capture the semantic features of the input data, enabling similarity-based operations. This concept is the backbone of this second type of employed method.

As with the zero-shot methodology, this one is used to frame articles into the political bias, reliability, and objectivity axes. The methodology is originally inspired by Peng et al. (2023), which used embedding-based retrieval and LLMs for efficient extraction of structured agricultural information from unstructured text sources. The main goal was to address challenges in pest identification and general agricultural data extraction from sources like news, research papers, and websites. It proposed a zero-shot pipeline approach to automatically extract entities, attributes, and their descriptions, transforming them into structured data. This idea was later adapted into IndiVec (Lin et al., 2024), a framework for political bias detection in news articles.

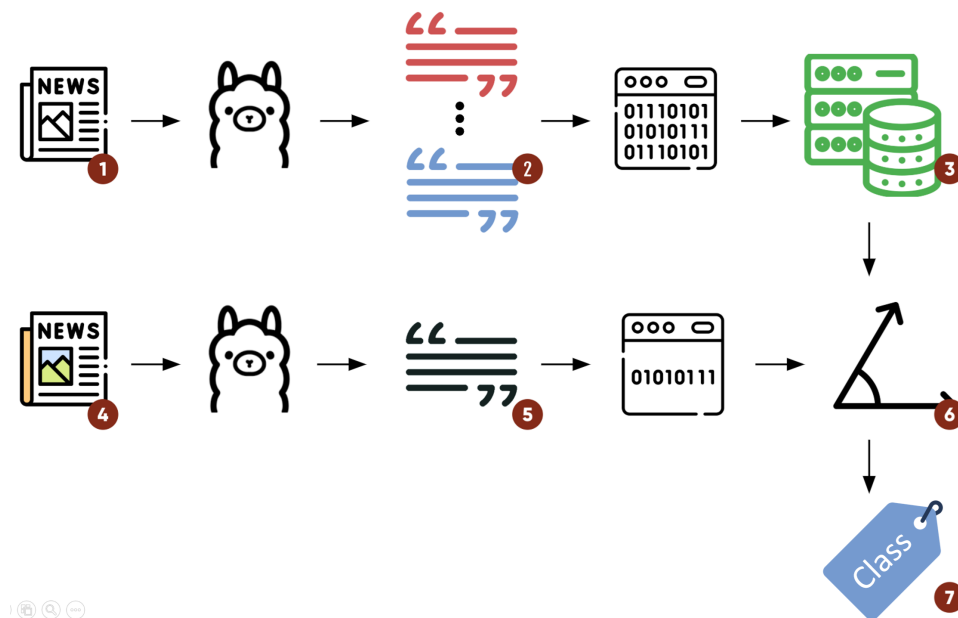


Figure 3.1: Vector databases classification pipeline

Figure 3.1 visually represents the vector database methodology for classifying news articles. The top row represents what could be considered the “training” of the system, although no training is involved, and the second row represents the testing/deployment phase. The process starts at step 1 where several labeled articles from the training set are provided to the LLM, one by one, through a series of prompts that also include the label and an instruction to generate an *indicator*, considering said label. The main idea is to guide the model to generate a small snippet of text that characterizes the article itself, based on the label. This label is attached to the generated indicator, as shown in step 2 of the schematic. To give an actual example, a right-leaning indicator of political bias generated by LLaMA 3.3 70B is “it refers to immigration as something negative

for society and associates it with problems of security and integration”. Regarding the prompts, this methodology has another peculiarity when compared to the first one: prompts are different depending not only on the axis but also on the label itself. This is to really enforce the label into the model to limit its chances of diverging from what is asked. The prompt only mentions one side of the problem, the one that regards the label of the article at stake, completely ignoring the fact that there are articles with the opposite label. To be consistent with the previous section, the following prompt refers to the system prompt used for objectivity classification, and specifically when the label is “objective”:

You are an assistant specialized in assessing the objectivity of articles based on their language, structure, and transparency. Your task is to identify indicators of objectivity and present the response in a well-defined JSON format.

Guidelines:

*- Consider **only formal elements** of the text (language, terminology, tone, structure).*

Do not assess the truthfulness of the content.

- Identify only one indicator at a time.

- The answer should have the following format: {"objective": "description of the identified indicator"}

When you receive an article, analyze its form and extract just one indicator of objectivity, strictly following the specified format.

In terms of the user prompt, similar to before, it now provides detailed instructions for the indicator generation task and definitions to guide the model. This time, in a similar fashion to few-shot approaches, examples are added not only to show the model what an indicator should look like in size and structure, but also to reinforce the output format. Once again, to maintain a balanced space utilization, only a small portion of the prompts is shown here. All other prompts, including for “subjective” articles, as well as for the other two axes, are present in Appendix B.2. The following is an English translation of the user prompt used in the same objectivity classification task, and when the label is “objective”:

An objectivity indicator is defined here as a descriptive label that represents the presence of elements in an article that contribute to its impartiality and rigor.

What you will detect are specifically linguistic, structural, or stylistic signs that reinforce the perceived objectivity of the article.

The output should always be a JSON with the following format:

{"objective": "description of the identified indicator"}

Examples:

{"objective": "Presents concrete and verifiable data to substantiate the information."}

{"objective": "Avoids emotional language or subjective adjectives."}

{"objective": "Uses a neutral and descriptive tone, without expressing opinion."}

{"objective": "Provides multiple perspectives on the topic, without favoring a specific

position."}
{"objective": "Includes references to credible and verifiable sources."}
{"objective": "Clear and logical structure, facilitating the understanding of the information."}

*Assuming the following text as objective, extract **only one indicator** of objectivity from the text and format the output **exactly as in the examples**:*
 {ARTICLE BODY}

After this major step, every indicator is converted into an embedding vector and stored in a vector database, as well as its corresponding label, as it is denoted in step 3 of Figure 3.1. The conversion is done using a Sentence Transformer (Reimers and Gurevych, 2019), specifically paraphrase-multilingual-mpnet-base-v2, which performs well for similarity search tasks in a multilingual context. At this stage, every article has yielded a single indicator that is stored in vector format in the vector database, associated with the respective label of the article.

In step 4, the image represents the classification of a new instance of news. The process is initially very similar, and the LLM is prompted to generate a *descriptor*. These differ from indicators in the sense that descriptors are not associated with any ground truth label since they come from new, unlabeled data (or a test set). Nevertheless, descriptors should look very similar to indicators in any other way, as step 5 illustrates. The prompt itself is a neutral version of the prompts used to generate the indicators: it does not include a specific label for the article, uses examples from both labels, and generally a more neutral language. Once more, the prompts to generate descriptors are presented in Appendix B.2. Similarly to before, the newly generated descriptor is converted into a vector representation using the same embedding model and then matched against the whole vector database. Using a measure such as cosine similarity, as denoted in step 6, the top K most similar indicator vectors are chosen and used to classify the descriptor into one of the classes (step 7). This whole process is repeated for each of the three axes, creating three different vector databases.

One of the goals of this dissertation is also to develop an interactive visualization tool to plot the four axes in space. To do this, however, the classification has to be translated into a specific figure that can be represented on a numerical axis. For each of the three vector databases already built, a score of -100 or 100 is also attached to each of the stored indicators, based on the label. For political bias, -100 was assigned to each left-wing indicator and 100 to each right-wing indicator; for reliability, -100 to indicators representing fake-news and 100 otherwise; and for objectivity, -100 to subjectivity-related indicators and 100 to objectivity-related ones. As explained in the previous paragraph, the process involves a similarity search across the respective vector database, which retrieves a cosine similarity value for each indicator. Subsequently, a weighted average score is computed, where the cosine similarities between the descriptor under classification and the indicators serve as the weighting factors, which is expressed in Equation 3.1. This weighted average score necessarily falls within $[-100, 100]$, and then, depending on whether it is above or

below 0, the respective class is assigned for evaluation purposes.

$$\text{Axis Score} = \frac{\sum_{i=1}^n \text{similarity}_i \times \text{score}_i}{\sum_{i=1}^n \text{similarity}_i} \quad (3.1)$$

The advantages of this methodology are clear. Firstly, it is adaptable and flexible since it is not fine-tuning a model into a specific dataset, risking domain overfitting, but actually extracting the causes of bias within the text. This might be of significant help when working with a proxy dataset such as the one proposed for the political bias axis, as hinted at before. Secondly, it is an explainable method, contrary to deep learning strategies, since indicators and descriptors are represented in natural language, allowing for direct human interpretation. Third, like other prompt-based techniques, it doesn't require an enormous amount of data in order to build a sufficiently big vector database that would allow for an effective system. And finally, the fact that it can conveniently map each article into an axis value is also very useful in this context. This final value is what determines the axis score of the classified descriptor (and article).

3.3.3 Readability Assessment

The last framing axis represents the readability factor of a news article. Given that there are no labeled datasets for the task in Portuguese, an alternative methodology was needed. In this case, the solution is based on applying traditional NLP techniques from the same family as the ones from Section 2.2.1, and in particular, established metrics for readability assessment, in line with what [Rodrigues \(2024\)](#) proposed. These metrics are usually count-based and focus on word complexity and overall text complexity as measured by word length, sentence length, occurrence of pre-defined complex words, and other metrics. This approach adds the advantage of being very direct, highly interpretable, and computationally very cheap since there is effectively no machine learning technique involved.

The main issue with established readability metrics is that most of them were developed targeting the English language. Fortunately, [Carvalho de Lima Moreno et al. \(2022\)](#) used multiple linear regression analysis to adapt several common metrics to the Portuguese language, although for its Brazilian variant. Nevertheless, it should be a suitable adaptation to complete the task.

Flesch Reading Ease One of the earliest readability assessment metrics is the Flesch Reading Ease ([Flesch, 1948](#)), tracing all the way back to the 1940s. It was originally normalized for a scale that goes from 0 (poor readability) to 100 (high readability), and the formula for the Portuguese-adapted version is as follows:

$$\text{Flesch Reading Ease} = 227 - 1.04 \times \frac{\text{words}}{\text{sentences}} - 72 \times \frac{\text{syllables}}{\text{words}} \quad (3.2)$$

Automated Readability Index Introduced by [Smith and Senter \(1967\)](#), the Automated Readability Index (ARI) is also a popular solution. The authors were motivated to develop this method to provide an easy way to evaluate the readability of technical manuals and books from the U.S. Air Force. One of the advantages is the lack of need for word difficulty and sentence difficulty factors, since the formula includes only count-based factors. The adaptation for Portuguese from [Carvalho de Lima Moreno et al. \(2022\)](#) is defined as follows:

$$\text{Automated Readability Index} = 0.44 \times \frac{\text{words}}{\text{sentences}} + 4.6 \times \frac{\text{characters}}{\text{words}} - 20 \quad (3.3)$$

Flesch-Kincaid grade level As an adaptation from the original Flesch Reading Ease formula presented above, Kincaid presented in 1975 the Flesch-Kincaid formula ([Carvalho de Lima Moreno et al., 2022](#)). This formula has the objective of mapping the readability levels into corresponding years of education needed to understand the text, considering the U.S. education system. In its adaptation, the formula is:

$$\text{Flesch-Kincaid grade level} = 0.36 \times \frac{\text{words}}{\text{sentences}} + 10.4 \times \frac{\text{syllables}}{\text{words}} - 18 \quad (3.4)$$

Gulpease Index Lastly, the Gulpease index is based on Flesch Reading Ease but relies on characters rather than syllables, which is an easier metric to obtain. It was created by the University of Rome La Sapienza in 1987 and tailored from inception for the Italian language. Interestingly, and perhaps because Portuguese and Italian share the same Latin roots, [Carvalho de Lima Moreno et al. \(2022\)](#) did not find any statistically significant changes to the parameters proposed by the following original formulation:

$$\text{Gulpease Index} = 89 + 300 \times \frac{\text{sentences}}{\text{words}} - 10 \times \frac{\text{letters}}{\text{words}} \quad (3.5)$$

After having these four metrics defined, the next step is to put together a coherent readability measure that can be mapped into an axis. In this case, the interval defined was between 0 and 100, different from the -100 to 100 interval on the other three axes already explained. This is because readability, especially as defined here as the output of established metrics, does not seem to have a theoretical neutral point, as, for example, political bias. A value of 0 in the readability axis means the text is difficult to read, while a value of 100 means the opposite.

The readability measure proposed can be viewed as a percentile range, and the basis for calculating this percentile is the hard news articles from the MINT dataset. This choice is based on the fact that this portion of the dataset is composed of actual news articles from a varied range of sources with meaningful representation, as shown in Section 3.2.3. This combination of attributes makes it a good representation of the targeted domain. The process starts by evaluating all 6,000 articles from this subset using the four formulas described above, separately. This results in four

distributions of readability metrics and, to avoid any issue due to extreme outliers, every distribution is winsorized at the percentiles 1 and 99; after this, the distributions are stored¹. It is important to note at this stage that two of these established metrics, specifically the Automated Readability Index and Flesch-Kincaid grade level, are on an inverted scale compared to the other two metrics, which means that the smaller the output, the higher the readability level. After inverting these two metrics into the desired scale direction, all 4 metrics assess readability in a similar way. To assess the readability of a new, unseen article, the same four formulas are applied to the text body, and with the outputs retrieved, the four output values are matched against the stored results from MINT to retrieve the percentile at which this new article stands within MINT's distribution, for each corresponding formula. Finally, a simple average of the four percentiles obtained from this whole process is computed, and this is the final readability score with respect to the new article.

3.4 Prompt Optimization

During Section 2.2.4, several prompt-based methods were presented and, as a consequence, several ways to optimize LLM inputs into more desirable outputs. Optimization methods inherently depend on the type of prompting method employed and, in this dissertation, only discrete prompts are used for their simplicity, interpretability, and lack of need for parameter training of any kind. Nevertheless, there are very interesting ways to optimize discrete prompt usage that go beyond resampling several similar possibilities and simply choosing the best-performing option based on the performance on a separate validation dataset. A particularly elaborate method was proposed by Guo et al. (2023), and it avoids any gradient-based optimization by using evolutionary algorithms. The method makes use of the capabilities of LLMs to generate new candidate prompts considering standard evolutionary operators, effectively merging the strengths of both tools. This idea was adapted for this dissertation to develop a new method for prompt optimization, also based on a popular optimization algorithm.

3.4.1 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a population-based stochastic optimization technique inspired by the social behavior patterns observed in flocks of birds and schools of fish. Originally proposed by Kennedy and Eberhart (1995), PSO has since evolved into one of the most widely adopted metaheuristic algorithms for solving optimization problems across various domains.

At its core, PSO simulates the dynamics of a swarm of particles, where each particle represents a potential solution within the search space. These particles navigate the search space by updating their positions and velocities according to both their own experience and the experience of neighboring particles. The underlying principle is the exploitation of both individual cognition and social learning to locate optima more effectively.

¹This means 24,000 values stored (4 distributions of values for 6,000 articles each).

Each particle i has a position vector $\mathbf{x}_i \in \mathbb{R}^n$ and a velocity vector $\mathbf{v}_i \in \mathbb{R}^n$, where n is the dimensionality of the search space. The particle remembers its own best-known position $\mathbf{p}_i$ (personal best) and has access to the global best-known position $\mathbf{g}$ identified by any particle in the swarm. The standard PSO algorithm updates velocity and position using the following equations:

$$\mathbf{v}_i(t+1) = w\mathbf{v}_i(t) + c_1r_1(\mathbf{p}_i - \mathbf{x}_i(t)) + c_2r_2(\mathbf{g} - \mathbf{x}_i(t)) \quad (3.6)$$

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) + \mathbf{v}_i(t+1) \quad (3.7)$$

Labeling the remaining variables from equations 3.6 and 3.7, w is the inertia weight, controlling the impact of the previous velocity, c_1 and c_2 are cognitive and social acceleration coefficients, respectively, and r_1 and r_2 are random values uniformly distributed in $[0, 1]$.

PSO distinguishes itself from evolutionary algorithms by relying on relatively simple mathematical operations and few hyperparameters. It does not employ genetic operators such as crossover and mutation, thereby reducing computational complexity. Additionally, it brings the advantage of always keeping the same population size during the whole process, something that does not happen in the genetic algorithm applied by Guo et al. (2023).

3.4.2 Particle Swarm Prompt Optimization

At the end of their paper, Guo et al. (2023) left several suggestions for future work in the area. One of these recommendations was actually to try to implement prompt optimization using PSO. In this section, the adaptation of the original algorithm into a prompt optimizer is presented, specifically within the context of zero-shot classification.

First things first, the optimization process requires a dataset to assess the prompts' performance along the way, and a validation set per axis is separated beforehand for this purpose. To follow PSO's logic, each particle is represented by a prompt, which has memory of its own best position. Additionally, the best prompt from the whole run (global best) is also saved. The process starts with an initial set of simple, hand-made prompts to provide a good initialization to the algorithm. After this, the algorithm starts by looping through all prompts to perform zero-shot classification and evaluation (based on accuracy) on the validation set, so that all variables are initialized. The second step is another loop of n iterations, where all particles are cycled through one by one. At this stage, the particles are moved, that is, the prompts are changed. As in the original PSO algorithm, the position is changed based on the current position, the best position of the current prompt, and the best position of the whole population. Maintaining the same spirit of the whole dissertation, every prompt update is an operation performed by leveraging the capabilities of a large language model – LLaMA 3.3 70B in this case, because it has revealed to be both more capable and also cost efficient² than EuroLLM 9B. Right after the update, the new prompt has to

²LLaMA was run on the API from Deep Infra which is cheap comparatively speaking.

be evaluated and compared to the best version of the current particle, as well as the global best position. If the conditions are met, one or both of these variables are updated.

There are two types of prompt updates based on a specific condition that is checked at every iteration: whether the current prompt is or is not its best version yet. If the current prompt is not its best version, then both the best version and the global best are considered for the new prompt. To perform this operation, LLaMA is guided to merge the three particles using the following instructions:

Consider the following original prompt:

Prompt A: {CURRENT PROMPT}

Consider also the following prompt as an improved version of the first:

Prompt B: {BEST VERSION OF THE CURRENT PROMPT }

Lastly, consider also the following prompt as being the best prompt yet:

Prompt C: {GLOBAL BEST PROMPT}

Mutate Prompt A in a way it is {LEVEL 1} inspired by the characteristics of Prompt B and {LEVEL 2} inspired by the characteristics of Prompt C.

*The output should be **only** a JSON with the following format: {"prompt": "new Prompt A"}.*

This process is referred to as a complete update. The implementation allows for tuning the sensitivity of the update to the different inputs in an attempt to approximate this adaptation to the original PSO algorithm. This sensibility is controlled by the variables Level 1 and 2, and the user can choose between “low”, “medium”, or “high”, which changes the wording of the instructions passed down to the model and maps these variables to “lightly”, “moderately”, or “heavily”, in respective order.

The second type of prompt update happens when the second condition is met: the current prompt is its best version. In this case, the process is called a partial update, and the instruction is slightly changed to include only the global best prompt in the mutation. The instructions for the partial update, as well as the system prompts used in this mutation process, are all presented in Appendix B.3. The whole algorithm’s logic is condensed into pseudo-code format in Algorithm 1.

In summary, this particle swarm prompt optimization algorithm is the mixture of three main ideas: first, of course, the original PSO algorithm presented by [Kennedy and Eberhart \(1995\)](#); secondly, to leverage LLMs as natural language operators that are able to change and adapt text (prompts) upon request, allowing for an automated mutation system, which is something also done by [Guo et al. \(2023\)](#); and lastly, the use of multiple paraphrased prompts that are tested on a validation set, as done without the use of any optimization algorithm by [Jiang et al. \(2019\)](#) and [Haviv et al. \(2021\)](#).

Algorithm 1 Particle Swarm Prompt Optimization

```

1: Initialize global_best_prompt
2: Initialize global_best_score

3: for each prompt in {prompt1, prompt2, prompt3, prompt4, prompt5} do
4:   prompt.classify(validation_set)
5:   prompt.evaluate()

6:   if prompt.score > prompt.best_score then
7:     prompt.best ← prompt.current
8:     prompt.best_score ← prompt.score
9:   end if
10: end for

11: for iteration = 1 to n do
12:   for each prompt in {prompt1, prompt2, prompt3, prompt4, prompt5} do
13:     if prompt.current ≠ global_best_prompt then

14:       if prompt.current = prompt.best then
15:         prompt.mutation_single(global_best_prompt, level_1)
16:       else
17:         prompt.mutation_complete(global_best_prompt, level_1, level_2)
18:       end if

19:       prompt.classify(validation_set)
20:       prompt.evaluate()

21:       if prompt.score > prompt.best_score then
22:         prompt.best ← prompt.current
23:         prompt.best_score ← prompt.score
24:       end if

25:       if prompt.score > global_best_score then
26:         global_best_prompt ← prompt.best
27:         global_best_score ← prompt.score
28:       end if

29:     end if
30:   end for
31: end for

```

Chapter 4

Experimental Evaluation

Following the methodological exposition presented in the previous chapter, this chapter outlines the results of all experiments conducted throughout the dissertation. It begins with an evaluation of the classification system, providing a detailed analysis of its performance, including a section dedicated to the performance obtained when employing the proposed particle swarm prompt optimization algorithm. Next, noteworthy insights are compiled concerning the nature and characteristics of the indicators produced through the use of the vector database methodology. Following this, a section is dedicated to the readability axis which, although it cannot be assessed through the same lens as the other axes, still offers valuable insights when examined more closely. A discussion on the relationships between the axes is also included. Finally, the chapter concludes with practical observations and real-world considerations, highlighting the capabilities and implications of the overall classification system.

4.1 Results

Table 4.1 presents the results obtained in every experiment. The results are shown for the EuroLLM 9B model and for LLaMA 3.3 70B, each first with the zero-shot (ZS) approach and then when using the vector database (VD) methodology.

Table 4.1: Performance comparison across models for political bias, reliability, and objectivity. "ZS" refers to the zero-shot methodology; "VD" refers to the vector database approach. The F1-score metric presented is the weighted-averaged version.

		EuroLLM_{ZS}	EuroLLM_{VD}	LLaMA_{ZS}	LLaMA_{VD}
Political Bias	Accuracy	0.56	0.59	0.67	0.69
	F1	0.55	0.59	0.66	0.69
Reliability	Accuracy	0.66	0.80	0.87	0.84
	F1	0.63	0.79	0.87	0.84
Objectivity	Accuracy	0.61	0.69	0.89	0.84
	F1	0.55	0.68	0.89	0.84

The first and most immediate observation is that LLaMA consistently outperforms EuroLLM across all axes and evaluation methods. This aligns with prior literature suggesting that larger-scale models tend to demonstrate stronger capabilities, richer contextual understanding, and more nuanced reasoning skills compared to their smaller counterparts (Kaplan et al., 2020; Wei et al., 2022a). The scale advantage translates directly into improved classification accuracy and F1-score in every comparable instance. Quantitatively, the improvements brought by using LLaMA over EuroLLM vary by axis and methodology. When considering the same axis and methodology, performance gains range from as low as 5% to as high as 46%, with an average improvement of approximately 23.5% in accuracy. This demonstrates the substantial benefit of larger model sizes in bias detection tasks.

Among the three evaluation axes, political bias detection is consistently the most challenging, regardless of the model or methodology. Accuracy for this axis ranges from a low of 0.56 (EuroLLM_{ZS}) to a respectable 0.69 (LLaMA_{VD}); F1-scores follow a similar trend. This lower performance could be attributed to the subjective and context-dependent nature of political stance, which often requires implicit world knowledge, understanding of social and political dynamics, and cultural framing to classify accurately. Using the vector database approach, performance gains are modest, highlighting once more the inherent difficulty of the task.

Reliability classification appears to be more tractable than political bias detection. EuroLLM’s performance sees a notable jump from 0.66 to 0.80 in accuracy and from 0.63 to 0.79 in F1-score when shifting from zero-shot to VD-based classification. This 14-16% performance boost effectively brings the smaller model closer to LLaMA’s performance, which achieves 0.87 in both metrics in the zero-shot condition. Interestingly, for LLaMA, the VD methodology slightly reduces performance on this axis. This may hint at the strength of LLaMA’s inherent zero-shot reasoning capabilities, which may already be difficult to beat on this task. Conversely, the second method appears especially helpful for smaller models, which likely benefit more from external grounding information due to their limited parameter capacity.

Objectivity assessment follows a pattern similar to reliability but with slightly lower scores for EuroLLM and comparable results for LLaMA. The use of the VD improves EuroLLM’s performance significantly – accuracy increases from 0.61 to 0.69, and the F1-score from 0.55 to 0.68. This suggests once more that this method can effectively ground the model’s outputs in concrete examples, which is useful for distinguishing objective from subjective writing. LLaMA performs very strongly in the zero-shot setting, reaching 0.89 accuracy and F1, but again sees a slight decrease when using the VD approach. The drop to 0.84 in both metrics is not negligible and further underscores the model’s intrinsic ability for contextual interpretation.

The results suggest that for smaller models, such as EuroLLM, the VD approach is a clear and consistent upgrade, improving performance across all axes. Notably, reliability classification with VD reaches 80% accuracy – on par with the performance of LLaMA in a similar setting. This makes a compelling case for this alternative methodology when using small-sized models that might otherwise fall short in zero-shot settings. For larger models, particularly LLaMA 3.3 70B,

zero-shot capabilities are already highly effective, especially for the less subjective tasks of reliability and objectivity. In some cases, the VD approach degrades performance slightly.

Lastly, it is important to note that the methodologies employed here do not involve any kind of fine-tuning, thereby reducing the risk of overfitting to specific datasets or distributions – an issue associated with fine-tuning methods, as reported by [Lin et al. \(2024\)](#). The approach presented here is truly generalizable and should retain its performance even when applied to other datasets for the same tasks. Furthermore, and specifically for the VD approach, the level of interpretability and transparency is kept at a high level since both indicators and descriptors are represented in natural language, allowing users to understand the reason behind the classification.

4.2 Particle Swarm Prompt Optimization

This section presents the results obtained by the prompt optimization technique designed during the development of this dissertation and explained in Section 3.4. The technique is based on the popular Particle Swarm Optimization algorithm and adapted for natural language operations, leveraging large language models.

For the experiment itself, 5 simple prompts were initially set by hand to provide a good initialization. The algorithm was run for 5 iterations with a validation set of 100 articles, randomly retrieved from the dataset of each respective axis, ensuring class balance and that the articles are not present in the test sets used before. These figures are relatively low because of the amount of API requests the process requires – for each iteration and for each prompt, the whole validation dataset is classified entirely, making it costly to use the whole available data for the purpose. Additionally, this has to be done for the political bias, reliability, and objectivity axes. The sensitivity levels introduced in Section 3.4 were all set to “medium”. Figure 4.1 shows the accuracy of each prompt and its evolution at each iteration of the algorithm, specifically for the political bias detection problem.

From the figure, it can be seen that the performance of each prompt improves, in general, across the whole run. Every prompt finishes with a better accuracy than what it started with, and the best one finishes surprisingly high with an accuracy of over 80%. Table 4.2 presents the performance of the optimized prompt for each axis alongside that of the standard zero-shot approach (repeated here for convenience purposes), both evaluated on the exact same test sets to ensure a fair comparison. These optimized prompts are provided in Appendix B.3.

The results consistently demonstrate that the optimization methodology yields better or equivalent performance compared to the zero-shot approach. In the domain of political bias, the Particle Swarm Prompt Optimization (PSPO) approach improves both accuracy (from 0.67 to 0.71) and F1-score (from 0.66 to 0.70). Throughout the dissertation, political bias has been hinted at being the most challenging and nuanced axis. Due to this, it was also the one with the most upside, which was confirmed by a notable increase in performance. Also noteworthy is the difference in performance of the optimized prompt on the validation set and on the test set, highlighting the importance of a rigorous evaluation process.

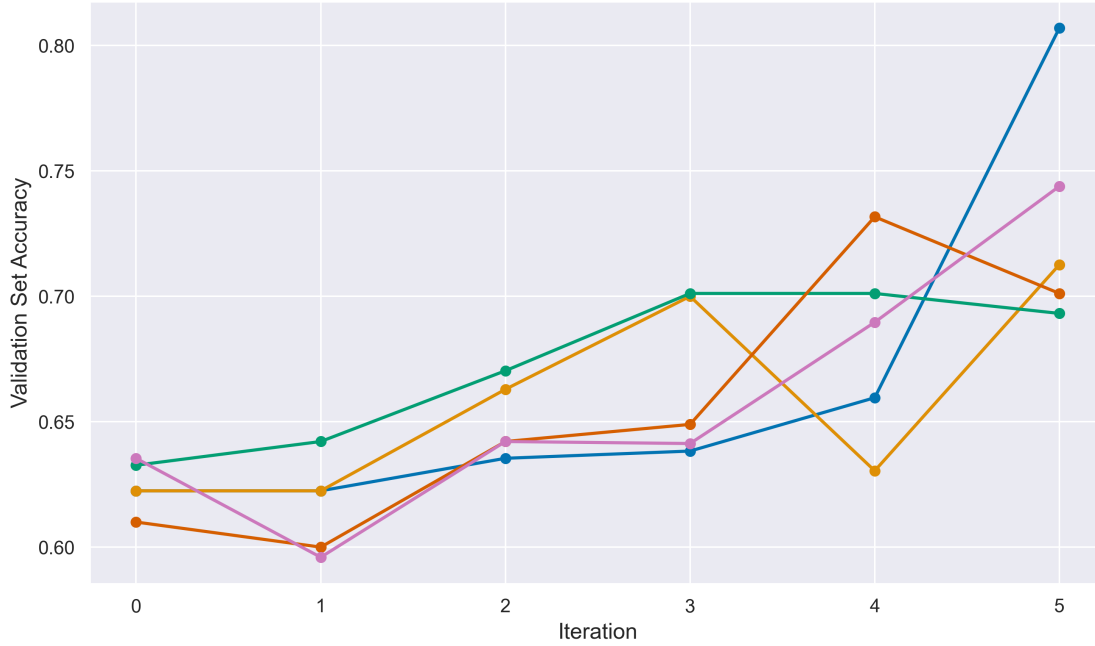


Figure 4.1: Optimization trajectory on political bias. Each line represents one particle (prompt).

Table 4.2: Performance comparison for political bias, reliability, and objectivity. "ZS" refers to the zero-shot methodology; "PSPO" refers to the prompt-optimized alternative. The F1-score metric presented is the weighted-averaged version.

		LLaMA _{ZS}	LLaMA _{PSPO}
Political Bias	Accuracy	0.67	0.71
	F1	0.66	0.70
Reliability	Accuracy	0.87	0.89
	F1	0.87	0.88
Objectivity	Accuracy	0.89	0.90
	F1	0.89	0.89

Improvements in reliability and objectivity are more modest but still present. For reliability, both accuracy and F1-score show small gains under the optimized setup, increasing from 0.87 to 0.89 in accuracy and from 0.87 to 0.88 in F1-score. Objectivity results show minimal change, with both ZS and PSPO achieving an identical F1-score of 0.89, and accuracy increasing only marginally (from 0.89 to 0.90). This indicates that objectivity assessment may already be a relatively well-handled task under zero-shot conditions.

4.3 Answering the Research Questions

4.3.1 Research Question 1

The findings presented in Table 4.1 provide evidence to help answer the first research question on whether small- and medium-scale LLMs are capable of framing news articles in a media-bias detection context. To perform the task with a smaller model, such as EuroLLM 9B, has been shown to be a challenging fit in a standard zero-shot fashion. When enhanced with a retrieval-augmented strategy, such as the presented VD approach, the performance has increased significantly, proving that there are ways to improve the model's baseline performance. The increase in performance was particularly remarkable when framing articles on the reliability axis.

Regarding the medium-sized model – LLaMA 3.3 70B – it performs robustly in all tasks, even in the zero-shot setting, often outperforming its own VD-enhanced version. It reaches high performance in reliability and objectivity classification, indicating that medium-scale models already possess strong generalization and reasoning capabilities for media bias detection. Interestingly, for LLaMA, the VD approach sometimes slightly reduces performance, suggesting that its internal capacity for abstraction and contextual understanding may exceed the benefits offered by external retrieval.

In conclusion, the results indicate that small-scale LLMs provide modest performance on the task, which can be enhanced to an acceptable level under specific circumstances. On the other hand, medium-scale models are effective out-of-the-box in framing news content into bias-related categories. This finding highlights a promising direction for cost-effective and interpretable bias detection systems that do not rely on computationally expensive model fine-tuning.

4.3.2 Research Question 2

The results presented in this dissertation support a positive answer to the second research question. The adapted PSO-based technique consistently improved prompt performance across all three evaluated axes when compared to the standard zero-shot approach. The most significant gains were observed in the political bias task, which, being the most challenging, benefited the most from optimization, demonstrating the potential of PSO to fine-tune prompts in complex classification tasks.

Although improvements for reliability and objectivity were more modest, the optimized prompts still matched or exceeded zero-shot performance. These results collectively indicate that Particle Swarm Optimization, when adapted for prompt tuning with LLMs, is an effective methodology for improving classification accuracy. Therefore, the PSO adaptation presented in this research is shown to be a promising and well-performing strategy for prompt optimization in this context.

4.4 Indicator Analysis

A core component of the proposed classification system is the generation of natural language indicators – concise textual explanations that reflect how a given article corresponds to a particular label. These indicators serve a dual purpose: they provide transparency into model decisions and populate the underlying vector databases used in the retrieval process. This section presents a comprehensive analysis of the indicators generated by the classification system across all axes. By examining both the quantity and the spatial distribution of these indicators, the influence of model scale on the generation process and the quality of resulting vector embeddings is assessed. The first subsection focuses on general statistics, comparing the number of unique indicators produced by the two language models under identical conditions, revealing how larger models tend to generate more diverse and contextually nuanced indicators. The second subsection explores the structure of the resulting vector spaces, highlighting meaningful patterns in the embeddings, such as inter-class separability and the semantic richness of LLaMA’s outputs. Together, these analyses offer insights into how indicator diversity and nuance contribute to the overall performance of the system.

4.4.1 General Statistics

As outlined in previous chapters, the classification system that makes use of the VD methodology targets three out of the four axes studied in this dissertation, namely political bias, reliability, and objectivity – each axis built with a distinct dataset. The classification framework is structured such that each article contributes a single indicator to its respective vector database, and no duplicate indicators are allowed within each database. This deliberate decision makes it so that the total number of indicators generated by a model is not only a function of the size of the training data but also of the model’s capacity to extract and rephrase nuanced content into brief textual snippets that faithfully reflect the labeled attribute.

Table 4.3 presents the total number of unique indicators generated by the two models across the three classification axes. These results were derived under identical conditions: the same data splits, prompt structures, and post-processing for both models. This controlled setting allows for isolating and examining the models’ intrinsic differences when performing indicator generation.

Table 4.3: Number of unique generated indicators in each vector database

	EuroLLM 9B	LLaMA 3.3 70B
Political Bias	2,519	4,630
Reliability	900	7,505
Objectivity	522	9,234

The data reveals an interesting trend where, through all three axes, the larger LLaMA 3.3 70B model consistently generates a significantly higher number of distinct indicators compared to the EuroLLM 9B model. This difference is particularly pronounced in the reliability and objectivity

axes, where the 70B model produces around 8 and 18 times more unique indicators, respectively. These findings underscore the importance of model scale in tasks requiring contextualization and nuanced textual understanding. The increased parameter count appears to confer the 70B model with a stronger capacity to detect subtle semantic distinctions, resulting in a greater variety of output that better captures the heterogeneous nature of each news piece.

The EuroLLM 9B model, while capable of producing syntactically and semantically coherent indicators, exhibits a tendency to overfit to the style or structure of the examples provided in the prompt, resulting in repetitive and formulaic outputs, therefore repeating the same indicators many times for completely independent articles – which for itself does not mean the indicators do not suit the article. In contrast, the LLaMA 3.3 70B model possesses a richer semantic representation space, allowing it to differentiate between subtly different articles and translate these differences into equally nuanced indicators without merely replicating the examples provided in the prompt.

Interestingly, the political bias axis displays a significantly smaller gap between the two models, with LLaMA producing under 2 times the number of indicators produced by EuroLLM. This suggests that the nature of the axis itself influences the indicator generation variance. Political bias, as a more ideologically and culturally loaded concept, may lend itself to a broader spectrum of plausible interpretations even for smaller models.

Overall, these findings validate how model scale influences the richness and granularity of the results. From a system performance perspective, richer indicator sets contribute to more expressive and discriminative vector databases, enhancing the retrieval process and classification accuracy, which helps explain the performance gap between the two models under the VD setting reported in Section 4.1. In applications where nuance is critical – such as media bias detection – this added diversity is valuable. These results emphasize once again the trade-offs inherent in model selection: while smaller models may be computationally more efficient, they are also more limited.

4.4.2 Indicator Vector Space

Indicators are stored in their respective databases as vector embeddings to enable similarity search. These text embeddings, each with 768 dimensions, are generated for each indicator using a Sentence Transformer model. While this high dimensionality makes direct visualization impossible, there are effective techniques available to reduce dimensionality and enable meaningful plotting. One such technique is the t-distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton, 2008), a powerful dimensionality reduction technique commonly used to visualize high-dimensional data. It works by modeling the relationships between data points in a way that preserves local structure – meaning that similar items stay close together when mapped to a lower-dimensional space. Figure 4.2 shows the 2D representation of the indicators from each vector database, obtained by applying t-SNE.

The visualization yields very interesting observations that confirm previous conclusions. Firstly, the variety of outputs provided by the LLaMA model is clear by the increased cloud density of its respective vector databases, when compared to those of EuroLLM. Additionally, both plots relative to political bias are noticeably more muddled, with less segregation of classes, compared

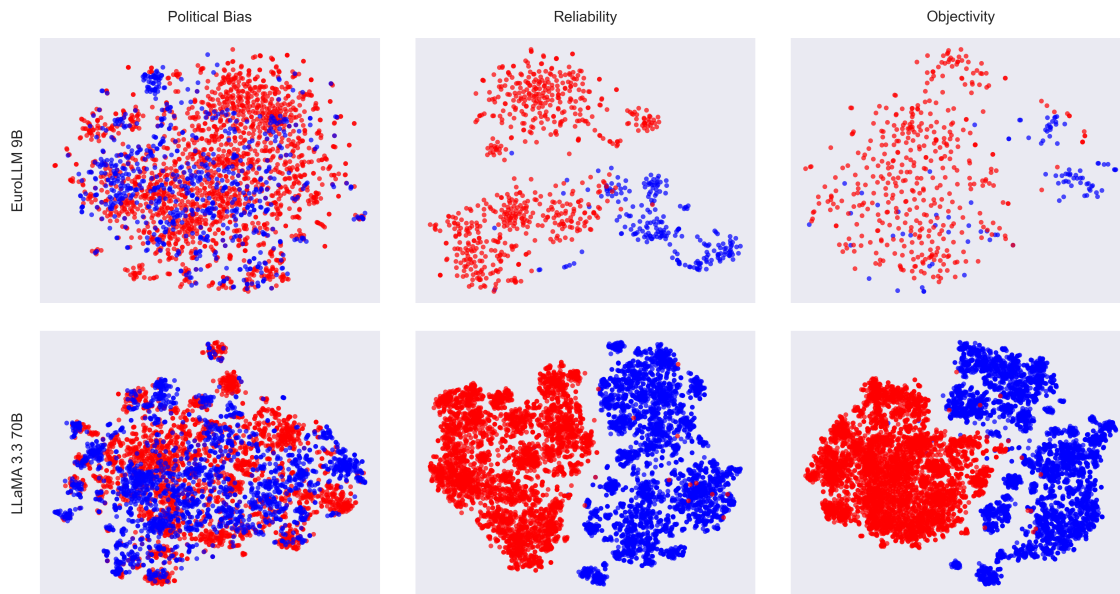


Figure 4.2: Indicator embeddings in 2 dimensions. The blue color is assigned to the classes “left”, “reliable”, and “subjective”, in the respective order of axes; red is assigned to the remaining classes.

to the reliability and objectivity counterparts. This hints once more at the increased subjectiveness of the topic of political stance, which seems to have a complexity not visualizable in two dimensions. The last and most remarkable observation is the very clear separation between classes in both reliability and objectivity indicators produced by LLaMA. The separation is so defined that a simple linear classifier, such as a logistic regression or a linear support vector machine, would achieve a high level of accuracy just by using the two dimensions obtained from the t-SNE process as features to predict the two classes.

4.5 Readability Assessment

Despite not having any labeled dataset and therefore not being suited for the same kind of evaluation as the other methodologies, the results from the readability assessment can also be explored.

The readability axis is assembled using 4 established readability metrics, adjusted for the Portuguese language. These metrics are based on various count-based measures from the text, such as the number of words, sentences, syllables, and characters, among others. In order to have a coherent classification process, the 6,000 articles from the hard news subset of the MINT dataset were scored for readability using the 4 metrics, essentially to serve as a base of comparison for new articles, as explained in detail in Section 3.3.3.

With the four distributions of readability scores obtained from this process, some observations can be made, with the first one represented in Figure 4.3. The figure shows a heatmap with the Pearson correlations between the scores obtained from each metric, namely the Automated

Readability Index (ARI), the Flesch-Kincaid grade level (FK), the Flesch Reading Ease (FRE), and the Gulpease Index (GI). To note, before computing the correlations, the results from both ARI and FK were inverted since their scoring is in the opposite direction of the remaining formulas. What can be immediately seen is that every correlation is extremely high, with a minimum of 0.91. This indicates two things: first, every metric is in agreement with each other which, for a well functioning system, is the expected behavior; and second, it indicates that using the four metrics might be redundant and choosing just one of them could be a viable approach for simplicity and efficiency.

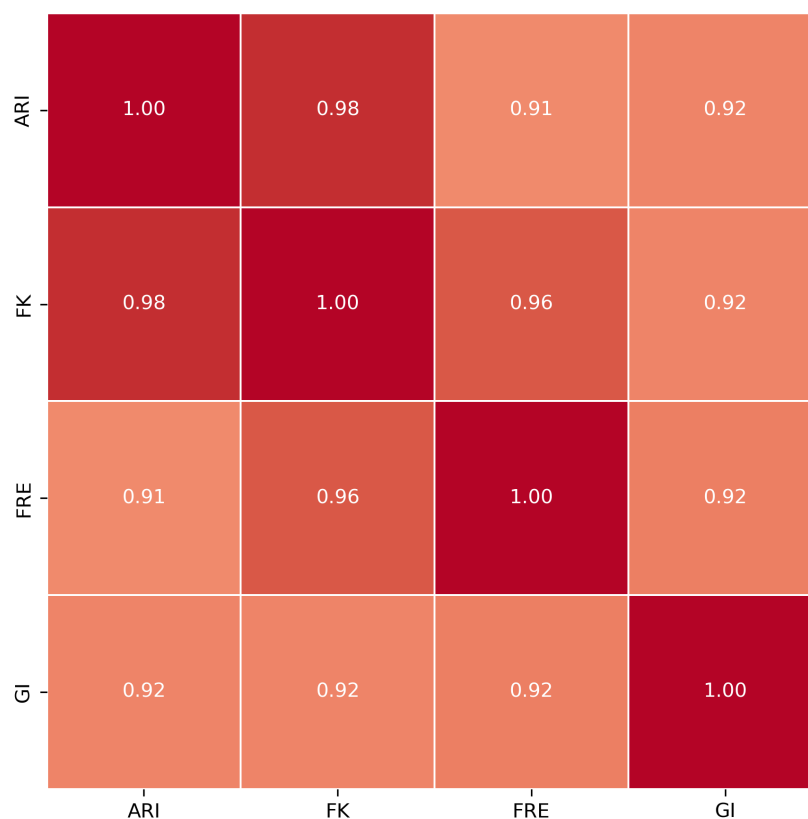


Figure 4.3: Correlation matrix for readability metrics

Another way of looking at the readability results is to directly consider their distributions – Figure 4.4 shows exactly this with a histogram for each metric. Interestingly, the readability scores of the articles from MINT seem to be close to normally distributed, which solidifies this subset of data as a good baseline of comparison for further assessment of new articles. Additionally, and in the same line with the observation from the previous paragraph, all four distributions seem very similar, with only the Gulpease Index one deviating slightly from the remaining.

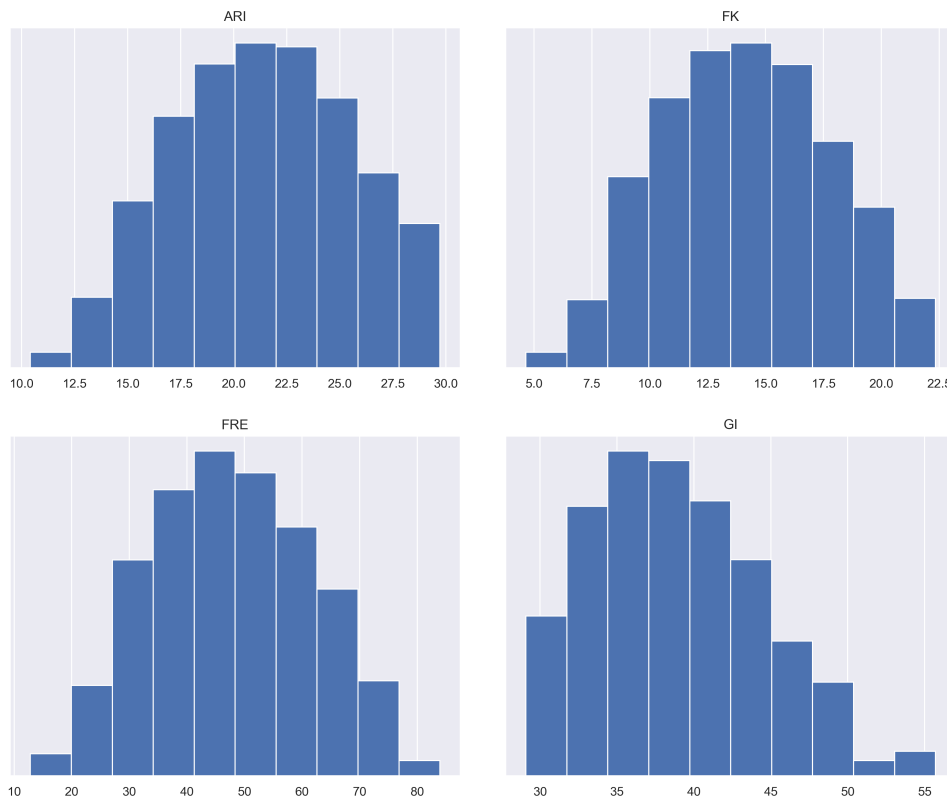


Figure 4.4: Readability scores distributions for every metric

4.6 Axes Relationships

To gain a deeper understanding of the four classification axes explored in this dissertation, it is useful to examine the relationships between them. In order to do this, a mixed dataset was put together, composed of test set articles from MINT (hard news subset labeled as objective) and Fake News PT (only the articles labeled as fake). This results in a total of 2200 articles from 14 different sources, ranging from credible news to outright fake news content, providing a good range of topics and writing styles to be put against each other.

Naturally, this mixed dataset was not yet fully classified for every axis, since each individual dataset was only classified for the respective axis during the development and evaluation of the classification systems. The first part (MINT) was only classified for objectivity and readability, whereas the second part (Fake News PT) was only classified for reliability. Therefore, all articles were classified for the remaining axes using the vector database approach built with LLaMA 3.3 70B.

A very interesting insight retrieved from this data is the noticeable article-level linear relationship between objectivity and reliability. Figure 4.5 shows this linear relationship on a hexbin plot, a less noisy version of a scatter plot, which is useful to avoid plotting 2200 points. Adding a simple linear regression quantifies the linear behavior between the two variables with an observed 69.5% R^2 measure. This is expected, since objective news articles should contribute to their reli-

ability while subjective content should contribute to the opposite. A logical observation like this reinforces the capability of the classification systems, since a large chunk of the articles in this sample were not even part of the dataset used for development of the vector database itself – these are out-of-domain articles, i.e., MINT was not used to build the reliability vector database, and Fake News PT was not used for the objectivity vector database. Regarding the remaining axes, no other pairwise relationship was found to this extent, at least at the article level.

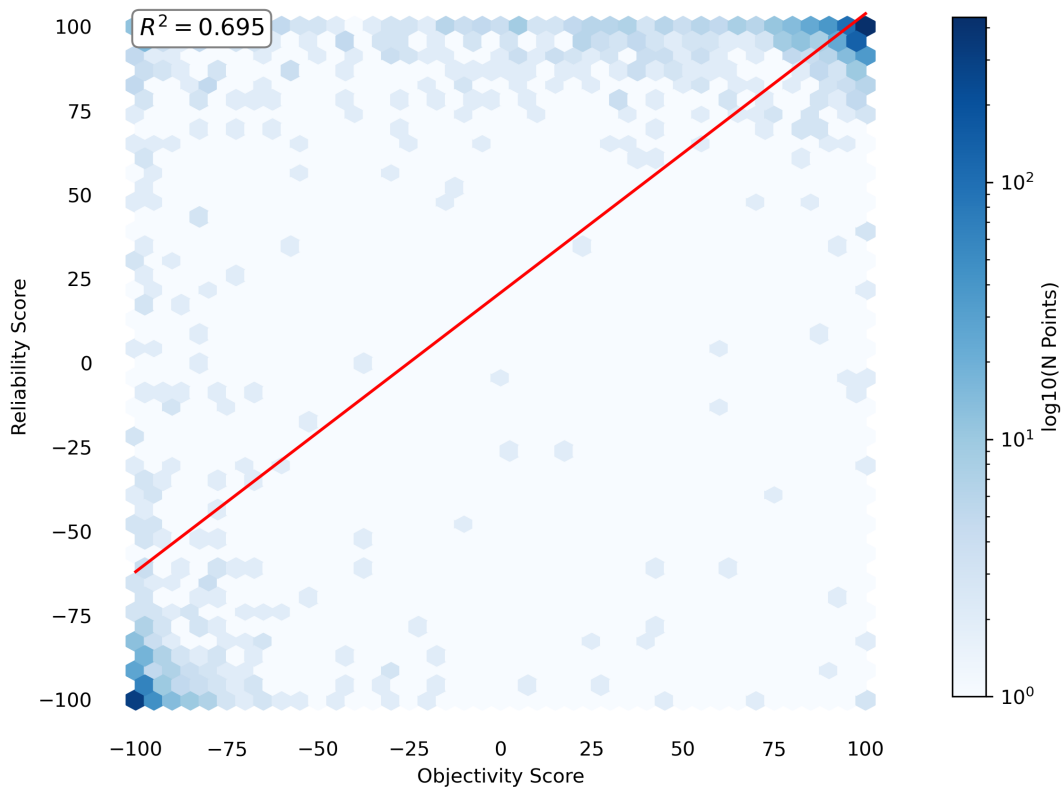


Figure 4.5: Relation between reliability and objectivity

A different way to approach the data is to aggregate everything by source, averaging out individual article scores. Given that there are 14 sources, the end results are four series of 14 average scores. Figure 4.6 shows a Pearson correlation matrix between every series. Although this is hardly enough data to draw definitive conclusions from, it is also very clear that there are signs of a strong linear correlation between the axes. At the source level, the relation between reliability and objectivity denoted above is kept very strong with a 0.98 coefficient. Furthermore, considering the meaning of political bias scores and the correlation obtained between the political bias axis and both reliability and objectivity, there is a tendency for right-wing biased sources to be considered less objective and less reliable, which is the same as saying that objective/reliable sources tend to be more left-wing biased than subjective/unreliable sources. As a final interesting remark on this analysis, the average readability score of a news source is hinted to be negatively correlated

with its reliability and objectivity, meaning that less reliable and more subjective sources tend to produce simpler, more readable content.

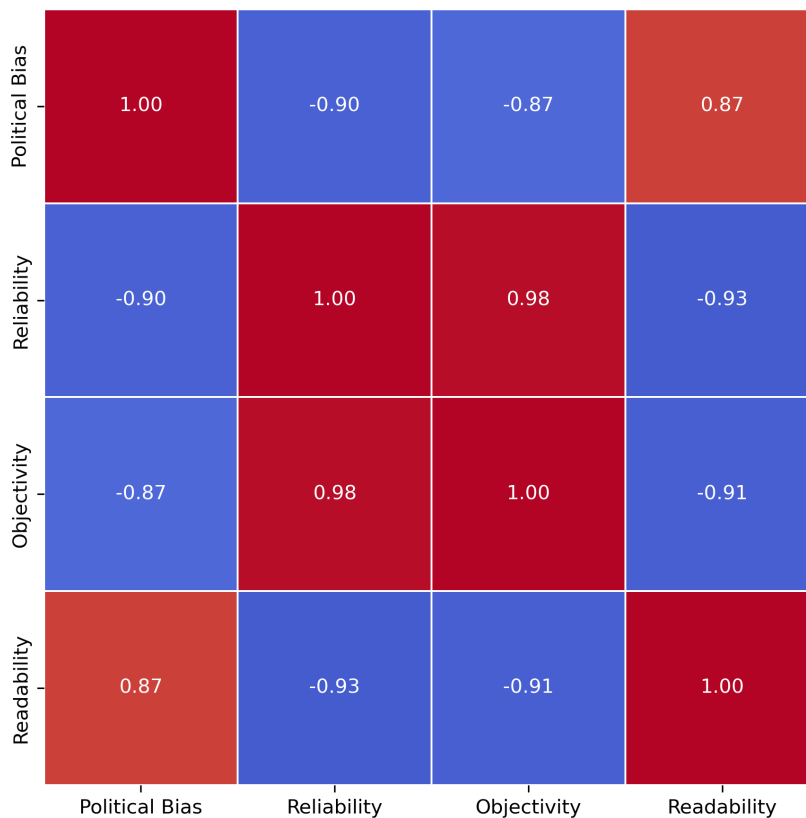


Figure 4.6: Pearson correlations between axes

4.7 Real World Takeaways

The last section of this chapter is the culmination of what the methods explored and tested during the whole chapter can actually be used for and how they can be helpful. Here, these methods are put in perspective to make actual real-world observations using the mixed dataset put together and explained in Section 4.6. The scores presented were computed using the vector database approach with LLaMA 3.3 70B and averaged out by article source, as shown in Table 4.4. The observations should always be considered through the lens of all the shortcomings highlighted across Chapters 3 and 4, as well as the reported system’s performance on unseen data from Section 4.1.

All media outlets categorized under the hard news subset from MINT exhibit, on the basis of their individual articles’ average, a slight left-wing bias. Within this group, Público, Diário de Notícias, and Expresso are the most left-leaning, whereas SIC Notícias, Lusa, and TSF demonstrate a comparatively lower bias. In contrast, the fake news subset skews to the right, with Direita Política being the most right-wing source – its ideological orientation aligning with its name,

Table 4.4: Average scores per source on each axis

	Political Bias	Reliability	Objectivity	Readability
Bombeiros 24	10	-7	-28	76
Direita Política	12	-73	-78	72
Magazine Lusa	10	-89	-91	85
Semanário Extra	7	-12	-1	64
Tuga Press	-7	-16	-32	72
Correio da Manhã	-14	82	63	49
Diário de Notícias	-25	86	51	52
Expresso	-18	86	55	50
Jornal de Notícias	-17	86	73	57
Jornal de Negócios	-18	94	79	51
Lusa	-14	89	52	46
Público	-26	88	58	43
SIC Notícias	-13	85	63	56
TSF	-14	86	44	49

which validates the system’s effectiveness in an interesting way. In terms of reliability, Jornal de Negócios emerges as the most trustworthy source in the dataset, achieving an average score of 94 on the -100 to 100 scale. The other expected reliable outlets fall in the 80s range, among which Correio da Manhã appears to be the least reliable. At the lower end of the reliability spectrum, Magazine Lusa and Direita Política register severely negative scores of -89 and -73, respectively. Following the findings discussed in Section 4.6, objectivity metrics align closely with reliability scores, reinforcing the same interpretative outcomes. When considering readability, Público stands out as the least accessible, scoring around 43. On the other hand, Jornal de Notícias ranks highest in readability among reliable sources with a score of 57. Interestingly, all outlets in the fake news category exhibit higher readability scores, which might pose an actual problem where unreliable news propagates more easily due to being more readable. The maximum is reached by Magazine Lusa with an average of 85, highlighting this inverse relationship between reliability and textual simplicity.

Chapter 5

Interactive Visualization Tool

The last goal of this dissertation was to build a visualization tool that is integrated with the whole classification system in a seamless fashion. One of the main objectives during the development of the interactive visualization tool, effectively showcasing the results of the automatic classification systems, was for it to be designed with user experience and flexibility in mind. The tool was built using Python’s Dash framework for the web interface and Plotly for dynamic data visualization. It was also deployed online using Hugging Face’s Spaces feature and is accessible through <https://huggingface.co/spaces/joao-oak/classificacao-media-pt>.

For the demonstration data, 200 articles were randomly chosen from the mixed dataset introduced in Section 4.6. The only requirement for the selection process was that the data should be evenly divided in terms of sources so that the data shown from the start would include a wide variety of types of news, spread over the whole range of the four axes.

After the first few tests, it was observed that some articles had an associated political bias descriptor noting that there was no sign of political bias – and understandably so. After the similarity search process and axis score computation, the result for these cases would often be slightly off to the right, instead of completely neutral as ideally it should be. To solve the issue at this deployment stage of the project, 15 neutral indicators were added manually to the political bias vector database with different versions of the sentence “*It does not show any political bias*”, covering a varied set of possible output descriptors when assessing a new example. Then, for the score calculation itself, if 3 of the neutral indicators are within the top 10 most similar, the final axis value given is 0 – an unbiased article. This is just one example of how the interpretability provided by this methodology is helpful, as emphasized multiple times throughout this dissertation.

In terms of the user interface, the visualization tool is organized into two main sections: a settings panel on the left and a visualization area on the right, as displayed in Figure 5.1. Since the target news articles are from Portuguese sources, and since the news titles, article bodies, and every indicator and descriptor are written in European Portuguese, it only made sense that the tool would be presented in the same language. One of the first added features of the tool is its support for customizable plot types, enabling users to choose between 1D, 2D, or 3D scatter plots. Depending on the chosen dimensionality, users can assign any of the four classification dimensions

– political bias, reliability, objectivity, and readability – to the X, Y, or Z axes. Each axis comes with individual range sliders that allow fine-grained filtering of data points based on minimum and maximum values, empowering users to focus on specific subsets of the data.

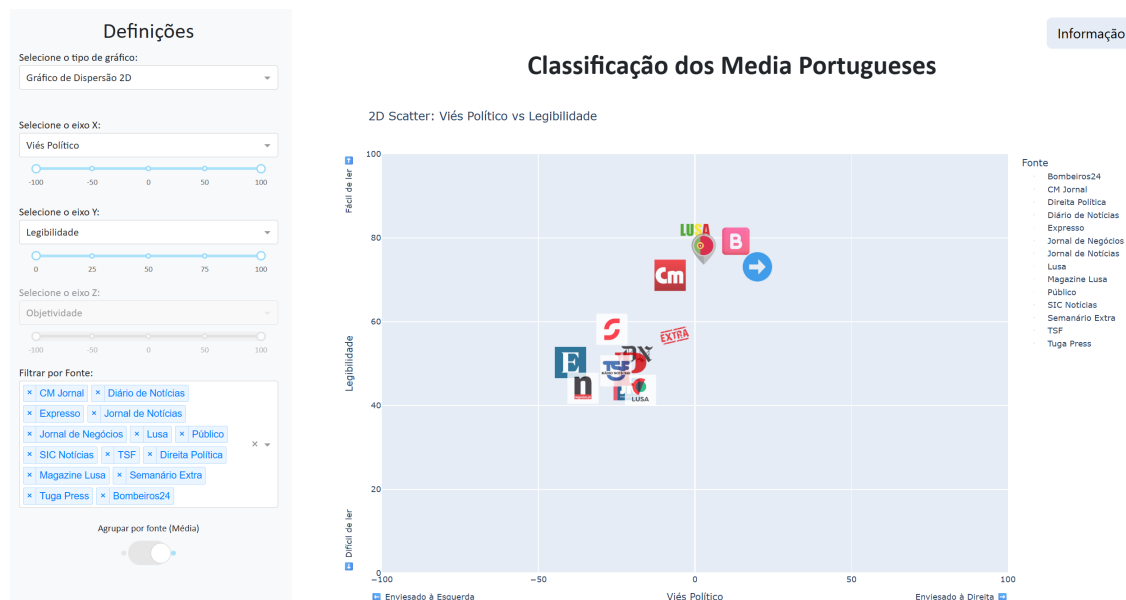


Figure 5.1: Tool UI with 2D plot using “Group By” functionality

To support comparative analysis and source-specific exploration, the tool includes a source filter that enables the user to isolate or exclude articles from specific media outlets. Furthermore, a grouping slider button allows users to toggle between viewing individual articles and a grouped-by-source visualization mode. When this mode is activated, each media outlet is represented by a single data point – the average of all its articles computed separately for each selected axis. These aggregated points are visually distinguished by the logo of the corresponding media outlet¹, offering a compact and visually pleasing summary of source-level trends and relationships.

Additionally, regardless of the kind of plot chosen, the tool shows interesting data when hovering the cursor on top of a specific article. An example of this is displayed on Figure 5.2. Firstly, the title of the specific article is written on top of the text box, and then the source is also specified. Adding up to that, the exact axis values are given and, most conveniently, the descriptors generated for that specific article that served as the basis for the classification are also displayed for transparency and to justify the axis value attributed. It must be said that this hover data is dynamically updated, meaning that it is dependent on the type of chart, as well as the specific axes chosen. In the case of Figure 5.2, this is an example taken from the 2D plot using political bias and reliability as the axes, and these are the values and descriptors shown. For the readability axis, only the axis value is available, as no descriptor is generated in that case. Another possible interaction is offered when clicking any of the data points displayed in the plot, which will open a pop-up window showcasing the full text body of the corresponding news article.

¹The logos feature is not supported for the 3D chart specifically, as it is interactive for rotation in every axis.

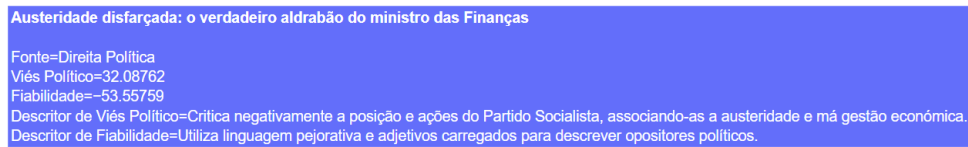


Figure 5.2: Hover data shown for one article

Beneath the main plot area lies the live classification module, a feature that enhances the exploratory nature of the tool. Users can dynamically classify new articles in two ways: either by uploading a `.txt` file or by providing a URL from a predefined set of supported Portuguese news sources (Expresso, Público, Eco Sapo, and Diário de Notícias). Naturally, if the article is provided through a URL, it should not be behind a paywall. This part of the UI is shown in Figure 5.3. The classification is performed in the exact same way as for the articles in the demo dataset, using the vector database approach built with LLaMA 3.3 70B with the addition of the politically-neutral indicators. Once classified, these new articles are immediately visualized in the chart and highlighted with a red cross to distinguish them from the pre-existing dataset. To note, the grouping functionality is deliberately disabled for live-classified articles, as they may originate from different sources and averaging them would not make much sense.

Arraste ou Seleccione um Artigo para Classificar (formato .txt)

Ou Insira um Link (Apenas para Expresso, Público, Eco Sapo, e Diário de Notícias)

<https://exemplo.com/artigo-noticias>

Submeter Link

Figure 5.3: Live classification functionality input

Lastly, in the upper right corner of the page, there is a clickable “Information” button that opens a pop-up window with a set of topics explaining the main functionalities of the tool, as well as the classification system from a very high level, targeting non-technical users for a broader reach. All features described are fully interoperable, meaning users can mix and match settings freely – choosing plot dimensions, filtering ranges, selecting sources, grouping data, and adding new articles – without disrupting the system’s functionality. This ensures a seamless and positive user experience, encouraging analysis and open-ended exploration. Appendix C includes additional figures showing different views of the tool, illustrating its functionalities.

In summary, the visualization tool was built not only to present the outputs of the classification system on demand, but to serve as an interactive exploratory environment where users can uncover patterns, compare sources, and do everything with minimal friction.

Chapter 6

Conclusion

This chapter serves as the closing point for this dissertation, wrapping up what was presented in the previous chapters by summarizing everything in general and from a high level, highlighting the main contributions of the dissertation, and finally leaving some suggestions for future research in the field.

6.1 Summary

Media bias is a very significant subject, and many news outlets, despite journalistic principles advocating for factuality and neutrality, often exhibit bias through framing, word choice, fact selection, and other means. This has several effects on public perception and policy decisions, resulting in declining trust in traditional forms of mainstream media.

The primary objective of this dissertation was to develop an automated system for framing media articles with some degree of explainability, allowing users to identify and understand the system's classifications. Articles are framed along four axes: political bias, reliability, objectivity, and readability. Furthermore, the classification is done by employing Large Language Models, such as EuroLLM 9B and LLaMA 3.3 70B, and prompting techniques to the underexplored domain of European Portuguese media. The last piece of the dissertation is an interactive visualization tool that incorporates the classification system and represents the media's landscape from several axes-related perspectives.

Regarding methodology in more detail, prompt-based methods are the main choice, covering three of the four classification axes. These methods consist of prompting LLMs to perform a desired task by leveraging their ability to model language probabilities. Two different prompt-based techniques are explored in this dissertation: zero-shot prompting and a method based on vector database techniques. On top of this, a new prompt optimization method is presented, based on the popular Particle Swarm Optimization algorithm, specifically targeted for zero-shot prompting

tasks. As for the remaining axis, readability is assessed by using established readability metrics applicable to the Portuguese language.

The results demonstrate that larger language models, such as LLaMA 3.3 70B, consistently outperform smaller models like EuroLLM 9B across bias detection tasks. However, the use of the vector database approach enhances the somewhat mild zero-shot performance of the smaller model. For LLaMA, the zero-shot performance is good in general, and the use of the vector database methodology occasionally leads to performance declines, suggesting that the model may already possess sufficient contextual understanding. Importantly, the methods employed avoid fine-tuning, thereby supporting greater generalizability and reducing the risk of overfitting. It is also shown that Particle Swarm Optimization works well as a prompt optimizer, generating marginal increases in LLaMA's zero-shot performance on every task.

Lastly, when looking deeper into the results, it was observable that reliability and objectivity have a strong linear relation, which is somewhat expected since both concepts are detected linguistically from similar signs and traits. From a higher level and with less statistical confidence, it was also observed that reliable media outlets are slightly biased toward left-leaning ideologies and produce less readable content in comparison to unreliable sources, which, on average, are politically right-wing biased and produce more accessible content in readability terms.

6.2 Contributions

This dissertation represents a meaningful contribution to the fields of media assessment, bias detection, and natural language processing, with a particular focus on addressing existing gaps in underexplored areas. The key contributions are summarized below.

A significant portion of this work is dedicated to addressing the lack of research and resources in media bias detection for the European Portuguese language. Unlike the wealth of studies focused on English-language media, particularly from the United States, Portuguese media have been largely overlooked. This dissertation tackles this gap by focusing on Portuguese media assessment, with an emphasis on identifying different characteristics beyond the commonly studied reliability or fake news detection. It also improves upon prior research by adopting an adaptable methodology that better handles proxy datasets and explores more representative and direct approaches to political stance detection.

By employing large language models and their in-context learning capabilities, this dissertation explores two distinct prompt-based methods for text classification tasks. These techniques, which have recently emerged in NLP research, are applied in the context of news articles classification, specifically from Portuguese media. This application not only extends the scope of prompt-based methods but also demonstrates their viability and effectiveness in addressing the challenges posed by lower-resourced languages like Portuguese, since they do not require a vast amount of data to be trained, and at the same time, offer a great deal of explainability into their classifications (in particular the vector database method presented).

The dissertation also includes an exploration of prompt optimization techniques for bias detection. In particular, a new optimization technique is introduced based on the Particle Swarm Optimization algorithm, adapted to natural language prompts and leveraging LLMs themselves as the optimization operators. By refining this algorithm for this specific use case, this work contributes to ongoing research into how prompts can be designed and fine-tuned for different NLP tasks.

Lastly, a custom-designed, interactive visualization tool was developed as part of the dissertation to provide users with a novel way to explore and analyze media articles. The tool represents media articles in a bias-representing space, offering flexibility and usability not commonly found in existing solutions. With its focus on axis variety and user-friendly interaction, the tool serves as a practical front-end, embedding the classification system into an actual application.

Together, these contributions position the dissertation as a valuable advancement in media assessment and bias detection, particularly in underexplored linguistic and cultural contexts, while leveraging state-of-the-art NLP methodologies and practical visualization tools to bridge the gap between academic research and deployment.

6.3 Future Work

Building upon the promising results and methodologies established in this dissertation, several paths for future research are worth pursuing.

Firstly, the Particle Swarm Optimization approach adapted for prompt optimization has shown encouraging results but still holds potential for refinement. Exploring different mutation techniques or considering velocity and position as two different variables, as in the original algorithm, could potentially result in more robust optimization outcomes. In parallel, it would be beneficial to experiment with alternative physics- or nature-inspired algorithms known for their performance in high-dimensional search spaces. These include, for example, Simulated Annealing, or even other swarm-based algorithms such as Ant Colony Optimization and Artificial Bee Colony. Comparing their effectiveness in prompt optimization scenarios could help identify the most suitable strategy for different prompt engineering tasks.

Another important line of research concerns data quality and availability, particularly for European Portuguese. The current lack of richly annotated datasets – especially in the domain of political bias – remains a major bottleneck. A dataset comprising professionally labeled news articles in Portuguese would not only enhance any method’s performance but also allow for more robust evaluation and conclusions.

Moreover, while this work has relied on established readability metrics, future studies could benefit from applying the same classification and evaluation methodologies to the readability axis itself. This would, however, necessitate the creation of another labeled dataset specific to Portuguese, where news texts would be annotated according to human-perceived readability levels.

Future efforts could also test the methodologies here presented on larger, more capable language models. Incorporating models such as GPT-4o, DeepSeek, or other frontier models, despite

hindering the ability to run the system locally, could provide deeper insights into the scalability and generalizability of the findings. Evaluating whether larger models exhibit improved sensitivity to prompt variations or bias detection nuances is an open and important question.

Expanding the analytical framework to include additional axes is also a plausible way to enrich this research. An attempt was made to add a sensationalism or clickbait factor axis which was cut during development for lack of data and poor performance from an alternative methodology. A successful attempt would contribute to a more complete palette of measures from which news could be framed, which is one of the most important contributions of this research.

Finally, from the specific point of view of this dissertation, a questionnaire could be done to assess the capabilities, user experience, visual appeal, and feature relevance of the interactive visualization tool presented here. This would allow for an external and independent evaluation of the tool coming from a large and representative pool of users.

Appendix A

Political Bias Dataset

Table A.1 details the authors considered to build the dataset for political bias detection, as well as the number of articles for each article. The articles came from a mix of sources including *Expresso*, *Público*, *Eco Sapo*, and *Dinheiro Vivo* – popular media outlets in Portugal that have a large library of opinion articles.

Table A.1: Authors considered in the political bias dataset

Author	Political affiliation	Number of articles
Daniel Oliveira	Left	1859
Carmo Afonso	Left	383
Isabel Moreira	Left	329
Ascenso Simões	Left	190
Pedro Adão e Silva	Left	99
Alexandra Leitão	Left	64
João Costa	Left	34
Sílvia Sousa	Left	14
Gonçalo Angeiras	Left	5
Mariana Mortágua	Left	3
Sofia Pereira	Left	3
Maria Escaja	Left	2
Marta Temido	Left	2
Henrique Raposo	Right	1897
Pedro Mexia	Right	868
José Miguel Júdice	Right	338
Pedro Gomes Sanches	Right	208
Henrique Brunay	Right	114
Sebastião Bugalho	Right	111
Luís Mira Amaral	Right	110
Miguel Poiares Maduro	Right	108
Óscar Afonso	Right	102
Carlos Guimarães Pinto	Right	45
Rodrigo Saraiva	Right	11
Miguel Morgado	Right	2

Appendix B

Prompts Utilized

The prompts presented across Chapter 3 were only a portion of the prompts employed during the development of this dissertation. This Appendix presents the remaining prompts. As a reminder, the news articles included in this study are written in European Portuguese, and so were the prompts. The prompts shown in this section were translated into English.

B.1 Zero-shot Prompts

B.1.1 Political Bias

System Prompt

You are a specialist in Portuguese politics. You are going to interpret text and identify political bias. In Portugal, Partido Socialista (PS), Bloco de Esquerda (BE), Livre, and Partido Comunista Português (PCP) are the left-wing parties. Partido Social Democrata (PSD), Iniciativa Liberal (IL), CDS, and Chega are the right-wing parties.

User Prompt

You are a model trained to identify the political bias of opinion articles or news reports. Your task is to classify the provided article as having either a "right-wing" or "left-wing" bias.

A text with a right-wing bias tends to emphasize values such as free market, reduction of the role of the state, conservatism, nationalism, and criticism of progressive policies.

A text with a left-wing bias tends to value state intervention in the economy, social justice and workers' rights, wealth redistribution, and criticism of neoliberalism.

The output must always be a JSON in the following format:

{"class": "right"} or {"class": "left"}

*The article below is **clearly biased**. Classify the bias of the following text and return*

only the correctly formatted response:

{ARTICLE BODY}

B.1.2 Reliability

System Prompt

You are an assistant specialized in evaluating the reliability of articles based on their language, structure, and transparency. Your task is to classify the reliability and present the response in a well-defined JSON format.

Guidelines:

- Consider only formal elements of the text (language, tone, structure, use of sources).

Do not assess the truthfulness of the content.

- The response must be in the following format: {"class": "given classification"}

When you receive an article, analyze its form and evaluate the level of reliability, strictly following the specified format.

User Prompt

You are a model trained to identify the reliability of opinion articles or news reports. Your task is to classify the provided article as either "reliable" or "unreliable", based on its language, structure, and transparency, without assessing the truthfulness of the content.

A reliable text tends to present verifiable sources, use neutral and objective language, have a clear and error-free structure, and avoid exaggeration and sensationalism.

An unreliable text tends to use sensationalist and emotionally charged language, contain grammatical or structural errors, and exaggerate or distort facts to support a point of view.

The output must always be a JSON in the following format:

{"class": "reliable"} or {"class": "unreliable"}

*The article below is **clearly classified** as one of these options. Classify the reliability of the following text and return **only the correctly formatted response**:*

{ARTICLE BODY}

B.2 Vector Database Prompts

B.2.1 Political Bias¹

User Prompt - Indicators (Right)

A political bias indicator is here defined as a descriptive label that represents the presence of political bias in a text.

¹The system prompt for political bias is exactly the same as before when performing zero-shot classification.

What you will detect is specifically right-wing political bias.

The output must always be a JSON in the following format:

{"bias": "description of the identified bias"}

Examples:

{"bias": "Refers to immigration as something negative for society."}

{"bias": "Mentions the advantages of low taxes for economic growth."}

{"bias": "Supports that the government should have little involvement in the economy and society."}

{"bias": "Devalues the welfare state and values individual merit."}

{"bias": "Criticizes the position of certain left-wing parties negatively."}

*Extract **only one new indicator** of right-wing political bias from the following text and format the output **exactly as in the examples**:*

{ARTICLE BODY}

User Prompt - Indicators (Left)

A political bias indicator is here defined as a descriptive label that represents the presence of political bias in a text.

What you will detect is specifically left-wing political bias.

The output must always be a JSON in the following format:

{"bias": "description of the identified bias"}

Examples:

{"bias": "Refers to immigration as something important for society."}

{"bias": "Supports government intervention on the economy and society."}

{"bias": "Criticizes capitalism."}

{"bias": "Speaks of climate change as an urgent problem."}

{"bias": "Criticizes the position of certain right-wing parties negatively."}

*Extract **only one new indicator** of left-wing political bias from the following text and format the output **exactly as in the examples**:*

{ARTICLE BODY}

User Prompt - Descriptors

A political bias indicator is here defined as a descriptive label that represents the presence of political bias in a text.

What you will detect is specifically political bias.

The output must always be a JSON in the following format:

{"bias": "description of the identified bias"}

Examples:

{"bias": "Refers to immigration as something negative for society."}

{"bias": "Supports government intervention on the economy and society."}

```

{"bias": "Mentions the advantages of low taxes for economic growth."}
{"bias": "Criticizes capitalism."}
{"bias": "Devalues the welfare state and values individual merit."}
{"bias": "Criticizes the position of certain right-wing parties negatively."}
Extract only one new indicator of political bias from the following text and format
the output exactly as in the examples:
{ARTICLE BODY}

```

B.2.2 Reliability

System Prompt - Indicators (Reliable)

You are an assistant specialized in assessing the reliability of articles based on their language, structure, and transparency. Your task is to identify indicators of reliability and present the response in a well-defined JSON format.

Guidelines:

- Consider **only formal elements** of the text (language, tone, structure, use of sources).
- Do not assess the truthfulness of the content.**
- Identify only one indicator at a time.
- The answer should have the following format: {"reliable": "description of the identified indicator"}

When you receive an article, analyze its form and extract just one indicator of reliability, strictly following the specified format.

User Prompt - Indicators (Reliable)

A reliability indicator is defined here as a descriptive label that represents the presence of elements in an article that may indicate greater credibility.

What you will detect are specifically linguistic, structural, or stylistic signs that reinforce the perceived reliability of the article.

The output should always be a JSON with the following format:

```

{"reliable": "description of the identified indicator"}

```

Examples:

```

{"reliable": "Presents verifiable sources for the information mentioned."}
{"reliable": "Uses neutral and objective language."}
{"reliable": "Avoids exaggerations or distortions when presenting the facts."}
{"reliable": "The text is well-structured and free of grammatical errors."}
{"reliable": "Provides multiple points of view on the topic addressed."}
{"reliable": "Presents statistics that support the argument."}

```

*Assuming the following text as reliable, extract **only one indicator** of reliability from*

*the text and format the output **exactly as in the examples:***

{ARTICLE BODY}

System Prompt - Indicators (Unreliable)

You are an assistant specialized in assessing the unreliability of articles based on their language, structure, and transparency. Your task is to identify indicators of unreliability and present the response in a well-defined JSON format.

Guidelines:

- Consider **only formal elements** of the text (language, tone, structure, use of sources). **Do not assess the truthfulness of the content.**
- Identify only one indicator at a time.
- The answer should have the following format: {"unreliable": "description of the identified indicator"}

When you receive an article, analyze its form and extract just one indicator of unreliability, strictly following the specified format.

User Prompt - Indicators (Unreliable)

An unreliability indicator is defined here as a descriptive label that represents the presence of elements in an article that may indicate less credibility.

What you will detect are specifically linguistic, structural, or stylistic signs that negatively impact the perceived reliability of the article.

The output should always be a JSON with the following format:

{"unreliable": "description of the identified indicator"}

Examples:

{"unreliable": "Uses sensationalist language to attract attention."}

{"unreliable": "Makes strong claims without citing verifiable sources."}

{"unreliable": "Contains grammatical and spelling errors."}

{"unreliable": "Includes exaggerations or distortions of facts to support a point of view."}

{"unreliable": "Uses vague and undefined terms like 'many experts say'."}

{"unreliable": "Excessive use of emotional language and loaded adjectives."}

*Assuming the following text as unreliable, extract **only one indicator** of unreliability from the text and format the output **exactly as in the examples:***

{ARTICLE BODY}

System Prompt - Descriptors

You are an assistant specialized in assessing the reliability of articles based on their language, structure, and transparency. Your task is to identify indicators of reliability

and present the response in a well-defined JSON format.

Guidelines:

*- Consider **only formal elements** of the text (language, tone, structure, use of sources).*

Do not assess the truthfulness of the content.

- Identify only one indicator at a time.

- The answer should have the following format: {"indicator": "description of the identified indicator"}

When you receive an article, analyze its form and extract just one indicator of the level of reliability, strictly following the specified format.

User Prompt - Descriptors

A reliability indicator is defined here as a descriptive label that represents the presence of elements in an article that may indicate higher or lower credibility.

What you will detect are specifically linguistic, structural, or stylistic signs that impact the perceived reliability of the article.

The output should always be a JSON with the following format:

{"indicator": "description of the identified indicator"}

Examples:

{"indicator": "Presents verifiable sources for the information mentioned."}

{"indicator": "Uses sensationalist language to attract attention."}

{"indicator": "Uses neutral and objective language."}

{"indicator": "Makes strong claims without citing verifiable sources."}

{"indicator": "Avoids exaggerations or distortions when presenting the facts."}

{"indicator": "Contains grammatical and spelling errors."}

{"indicator": "The text is well-structured and free of grammatical errors."}

{"indicator": "Excessive use of emotional language and loaded adjectives."}

*Extract **only one indicator** of the level of reliability from the text and format the output **exactly as in the examples**:*

{ARTICLE BODY}

B.2.3 Objectivity

System Prompt - Indicators (Subjective)

You are an assistant specialized in assessing the subjectivity of articles based on their language, structure, and transparency. Your task is to identify indicators of subjectivity and present the response in a well-defined JSON format.

Guidelines:

*- Consider **only formal elements** of the text (language, terminology, tone, structure).*

Do not assess the truthfulness of the content.

- Identify only one indicator at a time.
- The answer should have the following format: {"subjective": "description of the identified indicator"}

When you receive an article, analyze its form and extract just one indicator of subjectivity, strictly following the specified format.

User Prompt - Indicators (Subjective)

A subjectivity indicator is here defined as a descriptive label that represents the presence of elements in an article that indicate bias, opinion, or lack of factual rigor.

What you will detect are specifically linguistic, structural, or stylistic signs that reduce the perceived objectivity of the article.

The output should always be a JSON with the following format:

```
{"subjective": "description of the identified indicator"}
```

Examples:

```
{"subjective": "Uses opinionated language, expressing value judgments."}
```

```
{"subjective": "Shows explicit preference for one point of view without presenting counterpoints."}
```

```
{"subjective": "Includes assumptions or generalizations without support from verifiable data."}
```

```
{"subjective": "Makes frequent use of subjective adjectives and emotional language."}
```

```
{"subjective": "Presents persuasive arguments instead of neutral information."}
```

```
{"subjective": "Uses ambiguous or vague terms that make claims difficult to verify."}
```

Assuming the following text as subjective or opinion-based, extract **only one indicator** of subjectivity from the text and format the output **exactly as in the examples**:

{ARTICLE BODY}

System Prompt - Descriptors

You are an assistant specialized in assessing the objectivity of articles based on their language, structure, and transparency. Your task is to identify indicators of objectivity/subjectivity and present the response in a well-defined JSON format.

Guidelines:

- Consider **only formal elements** of the text (language, terminology, tone, structure). **Do not assess the truthfulness of the content.**
- Identify only one indicator at a time.
- The answer should have the following format: {"indicator": "description of the identified indicator"}

When you receive an article, analyze its form and extract just one indicator of objectivity/subjectivity, strictly following the specified format.

User Prompt - Descriptors

An objectivity indicator is here defined as a descriptive label that represents the presence of elements in an article that contribute to its greater or lesser impartiality and rigor.

What you will detect are specifically linguistic, structural, or stylistic signs that impact the perceived objectivity of the article.

The output should always be a JSON with the following format:

{"indicator": "description of the identified indicator"}

Examples:

{"indicator": "Presents concrete and verifiable data to substantiate the information."}

{"indicator": "Uses opinionated language, expressing value judgments."}

{"indicator": "Avoids emotional language or subjective adjectives."}

{"indicator": "Shows explicit preference for one point of view without presenting counterpoints."}

{"indicator": "Uses a neutral and descriptive tone, without expressing opinion."}

{"indicator": "Includes assumptions or generalizations without support from verifiable data."}

{"indicator": "Includes references to credible and verifiable sources."}

{"indicator": "Presents persuasive arguments instead of neutral information."}

*Extract **only one indicator** of objectivity from the text and format the output **exactly as in the examples**:*

{ARTICLE BODY}

B.3 Particle Swarm Prompt Optimization

System Prompt

*You are a specialist in prompt engineering. Your task is to improve, mutate, and generate a new prompt based on the inputs you will receive. The output should be **only** a JSON with the following format: {"prompt": "new generated prompt"}. Do not include any additional text other than the JSON.*

User Prompt - Partial Update

Consider the following original prompt:

Prompt A: {CURRENT PROMPT}

Consider also the following prompt as being the best prompt yet:

Prompt B: {GLOBAL BEST PROMPT}

Mutate Prompt A in a way it is {LEVEL 1} inspired by the characteristics of Prompt B.

*The output should be **only** a JSON with the following format: {"prompt": "new Prompt A"}.*

Optimized Prompts per Axis

Political Bias

Based on the content of the provided article, evaluate and classify the political bias as “left” or “right”, using the following JSON format: {"bias": "left"} or {"bias": "right"}.

Reliability

Evaluate the text in terms of the language used, organization, and clarity, also considering the level of transparency, classifying it as either {"class": "reliable"} or {"class": "unreliable"}, without considering the truthfulness of the information. Respond only with the classification.

Objectivity

Classify the following text as “objective” or “subjective”, based on the language, structure, and tone – without assessing its truthfulness. Respond only with a JSON: {"class": "objective"} or {"class": "subjective"}

Appendix C

Visualization Tool Views

This appendix displays a few examples of different possible views when interacting with the visualization tool. Every functionality can be mixed with each other so although the following images do not show every possible combination on functionalities, they are surely possible.



Figure C.1: 1D plot for political bias filtered for two sources

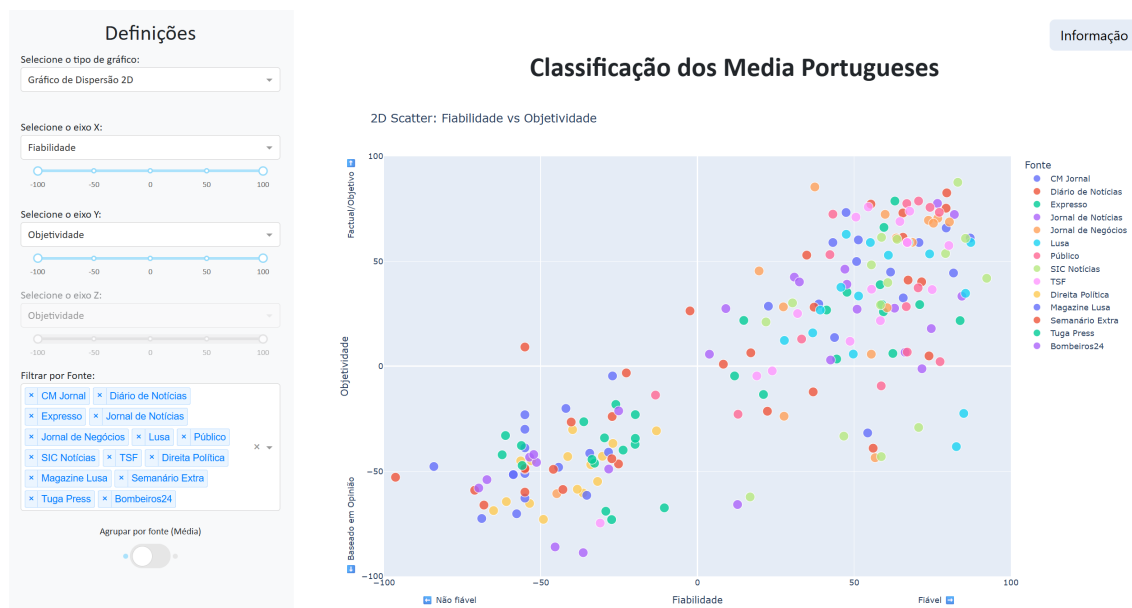


Figure C.2: Unfiltered 2D plot for reliability and objectivity

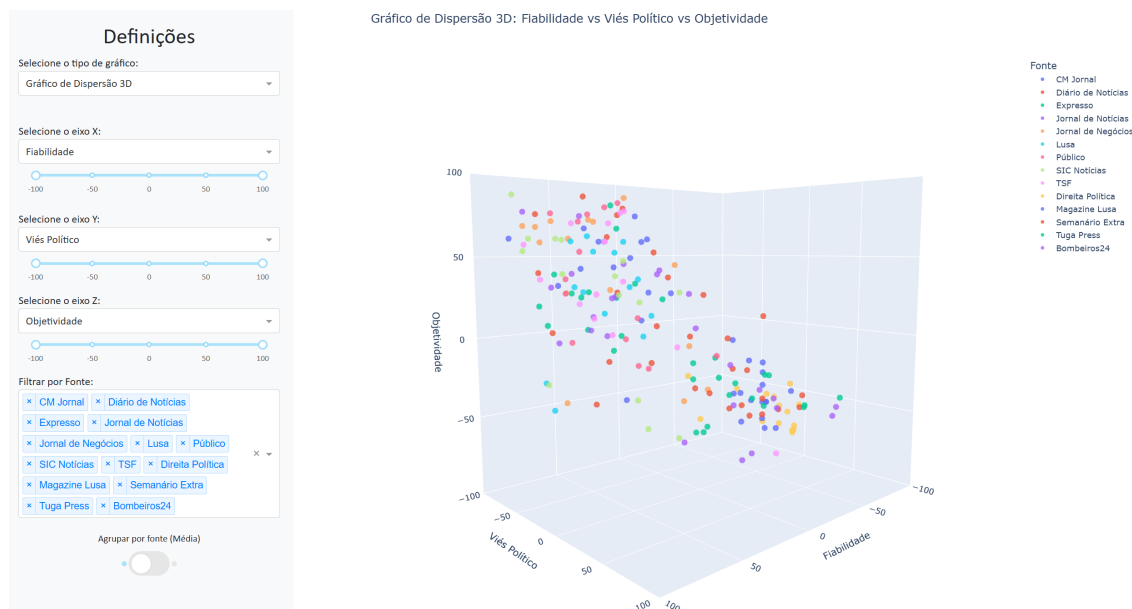


Figure C.3: Unfiltered 3D plot for reliability, political bias, and objectivity



Figure C.4: 1D plot for readability with source filters, grouping functionality, and live classification through URL

References

- COLING 1967 Volume 1: Conference Internationale Sur Le Traitement Automatique Des Langues*, 1967. URL <https://aclanthology.org/C67-1000/>.
- Hugo Queiroz Abonizio, Janaina Ignacio de Morais, Gabriel Marques Tavares, and Sylvio Barbon Junior. Language-independent fake news detection: English, portuguese, and spanish mutual features. *Future Internet*, 12(5):87, 2020. ISSN 1999-5903. URL <https://www.mdpi.com/1999-5903/12/5/87>.
- Ad Fontes Media. Interactive media bias chart, 2025. URL <https://app.adfontesmedia.com/chart/interactive>.
- R. Afonso and J. Rosas. Development of a smartphone application and chrome extension to detect fake news in english and european portuguese. *IEEE Latin America Transactions*, 22(4):294–303, 2024. ISSN 1548-0992. doi: 10.1109/TLA.2024.10472958. URL <https://github.com/ro-afonso/fake-news-pt-eu>.
- Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. Cross-domain fake news detection using a prompt-based approach. *Future Internet*, 16(8):286, 2024. ISSN 1999-5903. URL <https://www.mdpi.com/1999-5903/16/8/286>.
- AllSides. Media bias chart, 2024. URL <https://www.allsides.com/media-bias/media-bias-chart>.
- Talita Anthonio and Lennart Kloppenburg. *Team Kermit-the-frog at SemEval-2019 Task 4: Bias Detection Through Sentiment Analysis and Simple Linguistic Features*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019. doi: 10.18653/v1/S19-2177. URL <https://aclanthology.org/S19-2177/https://doi.org/10.18653/v1/S19-2177>.
- Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. Ext5: Towards extreme multi-task scaling for transfer learning, November 01, 2021 2021. URL <https://ui.adsabs.harvard.edu/abs/2021arXiv211110952A>.
- Larry Atkins. *Skewed: A Critical Thinker’s Guide to Media Bias*. Prometheus, 2016.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, September 01, 2014 2014. URL <https://ui.adsabs.harvard.edu/abs/2014arXiv1409.0473B>.
- Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. We can detect your bias: Predicting the political ideology of news articles, October 01, 2020 2020. URL <https://ui.adsabs.harvard.edu/abs/2020arXiv201005338B>.

- Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. *Testing and Comparing Computational Approaches for Identifying the Language of Framing in Political News*. Association for Computational Linguistics, Denver, Colorado, 2015. doi: 10.3115/v1/N15-1171. URL <https://aclanthology.org/N15-1171/https://doi.org/10.3115/v1/N15-1171>.
- Eric P. S. Baumer, Francesca Polletta, Nicole Pierski, and Geri K. Gay. A simple intervention to reduce framing effects in perceptions of global climate change. *Environmental Communication*, 11(3):289–310, 2017. ISSN 1752-4032. doi: 10.1080/17524032.2015.1084015. URL <https://doi.org/10.1080/17524032.2015.1084015>.
- Timothy Besley and Andrea Prat. Handcuffs for the grabbing hand? media capture and government accountability. *American Economic Review*, 96(3):720–736, 2006. doi: 10.1257/aer.96.3.720. URL <https://www.aeaweb.org/articles?id=10.1257/aer.96.3.720>.
- C. J. Beukeboom and C. F. Burgers. *Linguistic bias*, pages 1–19. Oxford Research Encyclopedias. Oxford University Press, 2017. doi: 10.1093/acrefore/9780190228613.013.439. URL <https://oxfordre.com/communication/>.
- Bimal Bhattarai. *A Rule-Based Framework for Interpretable Natural Language Processing*. Thesis, 2024.
- Biasly. Media bias charts, 2025. URL <https://www.biasly.com/media-bias-chart/>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, May 01, 2020 2020. URL <https://ui.adsabs.harvard.edu/abs/2020arXiv200514165B>.
- Hans-Jürgen Bucher and Peter Schumacher. The relevance of attention for selecting news content. an eye-tracking study on attention patterns in the reception of print and online media. 31 (3):347–368, 2006. doi: doi:10.1515/COMMUN.2006.022. URL <https://doi.org/10.1515/COMMUN.2006.022>.
- Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. Llms are few-shot in-context low-resource language learners, March 01, 2024 2024. URL <https://ui.adsabs.harvard.edu/abs/2024arXiv240316512C>.
- Y. E. Cakra and B. Distiawan Trisedya. Stock price prediction using linear regression based on sentiment analysis. In *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 147–154, 2015. doi: 10.1109/ICACSIS.2015.7415179.
- Danielle Caled, Paula Carvalho, and Mário J. Silva. Mint - mainstream and independent news text corpus. In Vlória Pinheiro, Pablo Gamallo, Raquel Amaro, Carolina Scarton, Fernando Batista, Diego Silva, Catarina Magro, and Hugo Pinto, editors, *Computational Processing of the Portuguese Language*, pages 26–36, Cham, 2022. Springer International Publishing. ISBN 978-3-030-98305-5.

- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. doi: doi:10.1126/science.aal4230. URL <https://www.science.org/doi/abs/10.1126/science.aal4230>.
- Dallas Card, Amber Boydston, Justin H Gross, Philip Resnik, and Noah A Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, 2015.
- Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. Assessing student explanations with large language models using fine-tuning and few-shot learning. *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 403–413. Association for Computational Linguistics, 2024. URL <https://aclanthology.org/2024.bea-1.33/>.
- Paula Carvalho, Danielle Caled, Mário J Silva, Bruno Martins, João Paulo Carvalho, Joaquim Carreira, João Pedro Fonseca, Teresa Gomes, and Pedro Camacho. Assessing news credibility: Misinformation content indicators. *Review*, Mar, 2021.
- Gleice Carvalho de Lima Moreno, Marco P. M. de Souza, Nelson Hein, and Adriana Kroenke Hein. Alt: um software para análise de legibilidade de textos em língua portuguesa, March 01, 2022 2022. URL <https://ui.adsabs.harvard.edu/abs/2022arXiv220312135C>.
- Li Chen, Guanliang Chen, and Feng Wang. Recommender systems based on user reviews: the state of the art. *User Modeling and User-Adapted Interaction*, 25(2):99–154, 2015. ISSN 1573-1391. doi: 10.1007/s11257-015-9155-5. URL <https://doi.org/10.1007/s11257-015-9155-5>.
- Sheng Chen, Akshay Soni, Aasish Pappu, and Yashar Mehdad. Doctag2vec: An embedding based multi-label learning approach for document tagging, July 01, 2017 2017. URL <https://ui.adsabs.harvard.edu/abs/2017arXiv170704596C>.
- Wei-Fan Chen, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. Learning to flip the bias of news headlines. In *International Conference on Natural Language Generation*, 2018.
- Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, December 01, 2014 2014. URL <https://ui.adsabs.harvard.edu/abs/2014arXiv1412.3555C>.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Un-supervised cross-lingual representation learning at scale, November 01, 2019 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv191102116C>.
- Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2(1):23, 2015. ISSN 2196-1115. doi: 10.1186/s40537-015-0029-9. URL <https://doi.org/10.1186/s40537-015-0029-9>.

- André Cruz, Gil Rocha, Rui Sousa-Silva, and Henrique Lopes Cardoso. *Team Fernando-Pessa at SemEval-2019 Task 4: Back to Basics in Hyperpartisan News Detection*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019. doi: 10.18653/v1/S19-2173. URL <https://aclanthology.org/S19-2173/https://doi.org/10.18653/v1/S19-2173>.
- André Ferreira Cruz, Gil Rocha, and Henrique Lopes Cardoso. On document representations for detection of biased news articles, 2020. URL <https://doi.org/10.1145/3341105.3374025>.
- Tanvi Dadu, Kartikey Pant, and Radhika Mamidi. Towards detection of subjective bias using contextualized word embeddings, February 01, 2020 2020. URL <https://ui.adsabs.harvard.edu/abs/2020arXiv200206644D>.
- Dave D'Alessio and Mike Allen. Media bias in presidential elections: A meta-analysis. *Journal of Communication*, 50(4):133–156, 2006. ISSN 0021-9916. doi: 10.1111/j.1460-2466.2000.tb02866.x. URL <https://doi.org/10.1111/j.1460-2466.2000.tb02866.x>.
- Alexander Dallmann, Florian Lemmerich, Daniel Zoller, and Andreas Hotho. Media bias in german online newspapers, 2015. URL <https://doi.org/10.1145/2700171.2791057>.
- Samantha D'Alonzo and Max Tegmark. Machine-learning media bias. *PLoS ONE*, 17:e0271947, 2022. doi: 10.1371/journal.pone.0271947. URL <https://ui.adsabs.harvard.edu/abs/2022PLoS..1771947D>.
- Gabriel Domingos De Arruda, Norton Trevisan Roman, and Ana Maria Monteiro. Analysing bias in political news. *JUCS - Journal of Universal Computer Science*, 26(2):173–199, 2020. ISSN 0948-695X 0948-6968. URL <https://doi.org/10.3897/jucs.2020.011>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, October 01, 2018 2018. URL <https://ui.adsabs.harvard.edu/abs/2018arXiv181004805D>.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, December 01, 2022 2022. URL <https://ui.adsabs.harvard.edu/abs/2023arXiv230100234D>.
- James N. Druckman and Michael Parkin. The impact of media bias: How editorial slant affects voters. *The Journal of Politics*, 67(4):1030–1049, 2005. doi: 10.1111/j.1468-2508.2005.00349.x. URL <https://www.journals.uchicago.edu/doi/abs/10.1111/j.1468-2508.2005.00349.x>.
- 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 01, 2019 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv190902670F>.
- Muhammad Fayaz, Atif Khan, Muhammad Bilal, and Sana Ullah Khan. Machine learning for fake news classification with optimal feature selection. *Soft Computing*, 26(16):7763–7771, 2022. ISSN 1432-7643.

- Lukas Feick, Karsten Donnay, and Katherine T. McCabe. The subconscious effect of subtle media bias on perceptions of terrorism. *American Politics Research*, 49(3):313–318, 2021. doi: 10.1177/1532673x20972105. URL <https://journals.sagepub.com/doi/abs/10.1177/1532673x20972105>.
- André Ferreira Cruz, Gil Rocha, and Henrique Lopes Cardoso. On sentence representations for propaganda detection: From handcrafted features to word embeddings. In Anna Feldman, Giovanni Da San Martino, Alberto Barrón-Cedeño, Chris Brew, Chris Leberknight, and Preslav Nakov, editors, *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 107–112, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5015. URL <https://aclanthology.org/D19-5015/>.
- Xavier Ferrer, Tom van Nuenen, Jose M. Such, and Natalia Criado. Discovering and categorising language biases in reddit, August 01, 2020 2020. URL <https://ui.adsabs.harvard.edu/abs/2020arXiv200802754F>.
- Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948. ISSN 1939-1854(Electronic),0021-9010(Print). doi: 10.1037/h0057532.
- Matthew Gentzkow and Jesse M. Shapiro. What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, 78(1):35–71, 2010. ISSN 0012-9682. doi: <https://doi.org/10.3982/ECTA7195>. URL <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA7195>.
- Matthew Gentzkow and Jesse M Shapiro. Media bias and reputation. *Journal of Political Economy*, 114(2):280–316, 2006. ISSN 0022-3808. doi: 10.1086/499414. URL <https://doi.org/10.1086/499414>.
- Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.
- Roei Golan, Sarah Ripps, Raghuram Reddy, Justin Loloi, Ari Bernstein, Zachary Connelly, Noa Golan, and Ranjith Ramasamy. Chatgpt’s ability to assess quality and readability of online medical information: Evidence from a cross-sectional study. *Cureus*, 15, 2023. doi: 10.7759/cureus.42214.
- Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. The voice of monetary policy. *American Economic Review*, 113(2):548–84, 2023. doi: 10.1257/aer.20220129. URL <https://www.aeaweb.org/articles?id=10.1257/aer.20220129>.
- Gregory Grefenstette, Yan Qu, James G. Shanahan, and David A. Evans. Coupling niche browsers and affect analysis for an opinion mining application, 2004.
- Tim Groeling. Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science*, 16(Volume 16, 2013):129–151, 2013. ISSN 1545-1577. doi: <https://doi.org/10.1146/annurev-polisci-040811-115123>. URL <https://www.annualreviews.org/content/journals/10.1146/annurev-polisci-040811-115123>.
- Tim Groseclose and Jeffrey Milyo. A measure of media bias*. *The Quarterly Journal of Economics*, 120(4):1191–1237, 2005. ISSN 0033-5533. doi: 10.1162/003355305775097542. URL <https://doi.org/10.1162/003355305775097542>.

- Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. Nela-gt-2022: A large multi-labelled news dataset for the study of misinformation in news articles, March 01, 2022 2022. URL <https://ui.adsabs.harvard.edu/abs/2022arXiv220305659G>.
- Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers, September 01, 2023 2023. URL <https://ui.adsabs.harvard.edu/abs/2023arXiv230908532G>.
- Felix Hamborg. *Towards Automated Frame Analysis : Natural Language Processing Techniques to Reveal Media Bias in News Articles*. Thesis, Universitat Konstanz, 2022.
- 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, 2019. ISSN 1432-1300. doi: 10.1007/s00799-018-0261-y. URL <https://doi.org/10.1007/s00799-018-0261-y>.
- T. U. Haque, N. N. Saber, and F. M. Shah. Sentiment analysis on large scale amazon product reviews. In *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, pages 1–6, 2018. doi: 10.1109/ICIRD.2018.8376299.
- Adi Haviv, Jonathan Berant, and Amir Globerson. Bertese: Learning to speak to bert, March 01, 2021 2021. URL <https://ui.adsabs.harvard.edu/abs/2021arXiv210305327H>.
- Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 06(02):107–116, 1998. doi: 10.1142/s0218488598000094. URL <https://www.worldscientific.com/doi/abs/10.1142/S0218488598000094>.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8): 1735–1780, 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Jiwoo Hong, Yejin Cho, Jaemin Jung, Jiyoung Han, and James Thorne. Disentangling structure and style: Political bias detection in news by inducing document hierarchy, April 01, 2023 2023. URL <https://ui.adsabs.harvard.edu/abs/2023arXiv230402247H>.
- Hairong Huo and Mizuho Iwaihara. Utilizing bert pretrained models with various fine-tune methods for subjectivity detection. *Web and Big Data*, pages 270–284. Springer International Publishing, 2020. ISBN 978-3-030-60290-1.
- C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, 2014. doi: 10.1609/icwsm.v8i1.14550. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know?, November 01, 2019 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv191112543J>.
- Zihang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Convbort: Improving bert with span-based dynamic convolution, August 01, 2020 2020. URL <https://ui.adsabs.harvard.edu/abs/2020arXiv200802496J>.

- Lalitha Kameswari, Dama Sravani, and Radhika Mamidi. *Enhancing Bias Detection in Political News Using Pragmatic Presupposition*. Association for Computational Linguistics, Online, 2020. doi: 10.18653/v1/2020.socialnlp-1.1. URL <https://aclanthology.org/2020.socialnlp-1.1/https://doi.org/10.18653/v1/2020.socialnlp-1.1>.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv e-prints*, art. arXiv:2001.08361, January 2020. doi: 10.48550/arXiv.2001.08361.
- J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pages 1942–1948 vol.4, 1995. doi: 10.1109/ICNN.1995.488968.
- Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. SemEval-2019 task 4: Hyperpartisan news detection. In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2145. URL <https://aclanthology.org/S19-2145/>.
- Michelle YoungJin Kim and Kristen Marie Johnson. *CLOSE: Contrastive Learning of Subframe Embeddings for Political Bias Classification of News Media*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022. URL <https://aclanthology.org/2022.coling-1.245/>.
- Anne C. Kroon, Damian Trilling, and Tamara Raats. Guilty by association: Using word embeddings to measure ethnic stereotypes in news coverage. *Journalism Mass Communication Quarterly*, 98(2):451–477, 2021. doi: 10.1177/1077699020932304. URL <https://journals.sagepub.com/doi/abs/10.1177/1077699020932304>.
- Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining, January 01, 2019 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv190107291L>.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, September 01, 2019 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv190911942L>.
- Susan Leavy. Uncovering gender bias in media coverage of politicians with machine learning, May 01, 2020 2020. URL <https://ui.adsabs.harvard.edu/abs/2020arXiv200507734L>.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, October 01, 2019 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv191013461L>.
- Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation, January 01, 2021 2021. URL <https://ui.adsabs.harvard.edu/abs/2021arXiv210100190L>.

- Luyang Lin, Lingzhi Wang, Xiaoyan Zhao, Jing Li, and Kam-Fai Wong. Indivec: An exploration of leveraging large language models for media bias detection with fine-grained bias indicators. *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1038–1050. Association for Computational Linguistics, 2024. URL <https://aclanthology.org/2024.findings-eacl.70/>.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):Article 195, 2023. ISSN 0360-0300. doi: 10.1145/3560815. URL <https://doi.org/10.1145/3560815>.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pre-training approach, July 01, 2019 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv190711692L>.
- Kevin Lo, Cornelius Joe, Samuel Philip, and Hidayaturrahman Hidayaturrahman. *Comparison of Machine Learning and Deep Learning Models for Detecting Cyberbullying*. 2024. doi: 10.1109/IVIT62102.2024.10692892.
- Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. Eurollm: Multilingual language models for europe, September 01, 2024 2024. URL <https://ui.adsabs.harvard.edu/abs/2024arXiv240916235M>.
- Jörg Matthes, Desirée Schmuck, and Christian von Sikorski. In the eye of the beholder: A case for the visual hostile media phenomenon. *Communication Research*, 50(7):879–903, 2023. doi: 10.1177/00936502211018596. URL <https://journals.sagepub.com/doi/abs/10.1177/00936502211018596>.
- Megan Brenan. Americans’ trust in media remains at trend low. *Gallup*, 2024. URL <https://news.gallup.com/poll/651977/americans-trust-media-remains-trend-low.aspx>.
- Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3, 2020. ISSN 2624-8212. doi: 10.3389/frai.2020.00055. URL <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2020.00055>.
- Rafael A. Monteiro, Roney L. S. Santos, Thiago A. S. Pardo, Tiago A. de Almeida, Evandro E. S. Ruiz, and Oto A. Vale. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. *Computational Processing of the Portuguese Language*, pages 324–334. Springer International Publishing, 2018. ISBN 978-3-319-99722-3.
- Ricardo Moura, Rui Sousa-Silva, and Henrique Lopes Cardoso. Automated fake news detection using computational forensic linguistics. In *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings 20*, pages 788–800. Springer, 2021. ISBN 3030862291.
- Sendhil Mullainathan and Andrei Shleifer. The market for news. *American Economic Review*, 95(4):1031–1053, 2005. doi: 10.1257/0002828054825619. URL <https://www.aeaweb.org/articles?id=10.1257/0002828054825619>.

- Przemysław Mućko. Sentiment analysis of csr disclosures in annual reports of eu companies. *Procedia Computer Science*, 192:3351–3359, 2021. ISSN 1877-0509. doi: <https://doi.org/10.1016/j.procs.2021.09.108>. URL <https://www.sciencedirect.com/science/article/pii/S1877050921018470>.
- Tarek Naous, Michael J. Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. Readme++: Benchmarking multilingual language models for multi-domain readability assessment, May 01, 2023 2023. URL <https://ui.adsabs.harvard.edu/abs/2023arXiv230514463N>.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 01, 2022 2022. URL <https://ui.adsabs.harvard.edu/abs/2022arXiv2203021550>.
- Zizi Papacharissi and Maria de Fatima Oliveira. News frames terrorism: A comparative analysis of frames employed in terrorism coverage in u.s. and u.k. newspapers. *The International Journal of Press/Politics*, 13(1):52–74, 2008. doi: 10.1177/1940161207312676. URL <https://journals.sagepub.com/doi/abs/10.1177/1940161207312676>.
- Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. Bias in word embeddings, 2020. URL <https://doi.org/10.1145/3351095.3372843>.
- Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. Newscube: delivering multiple aspects of news to mitigate media bias, 2009. URL <https://doi.org/10.1145/1518701.1518772>.
- Ruoling Peng, Kang Liu, Po Yang, Zhipeng Yuan, and Shunbao Li. Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data, August 01, 2023 2023. URL <https://ui.adsabs.harvard.edu/abs/2023arXiv230803107P>.
- Yilang Peng. Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, 68(5):920–941, 2018. ISSN 0021-9916. doi: 10.1093/joc/jqy041. URL <https://doi.org/10.1093/joc/jqy041>.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases?, September 01, 2019 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv190901066P>.
- Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. Automatically neutralizing subjective bias in text, November 01, 2019 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv191109709P>.
- Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news, August 01, 2017 2017. URL <https://ui.adsabs.harvard.edu/abs/2017arXiv170807104P>.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver?, February 01, 2023 2023. URL <https://ui.adsabs.harvard.edu/abs/2023arXiv230206476Q>.

- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, October 01, 2019 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv191010683R>.
- T. S. Rao and C. Andrade. The mmr vaccine and autism: Sensation, refutation, retraction, and fraud. (1998-3794 (Electronic)), 2011.
- Shaina Raza, Deepak John Reji, and Chen Ding. Dbias: detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics*, 17(1):39–59, 2024. ISSN 2364-4168. doi: 10.1007/s41060-022-00359-4. URL <https://doi.org/10.1007/s41060-022-00359-4>.
- Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 1650–1659, 2013.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL <http://arxiv.org/abs/1908.10084>.
- Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641, 2024. ISSN 0957-4174. doi: <https://doi.org/10.1016/j.eswa.2023.121641>. URL <https://www.sciencedirect.com/science/article/pii/S0957417423021437>.
- José Frederico Gomes Ferreira Marques Rodrigues. *Readability Assessment and Text Simplification through Open-Source Large Language Models*. Thesis, Universidade do Porto, 2024.
- 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-*, May 01, 2023 2023. URL <https://ui.adsabs.harvard.edu/abs/2023arXiv230506721R>.
- João Filipe Carriço Rodrigues. *Fake News Classification in European Portuguese Language*. Thesis, Instituto Universitário de Lisboa, 2020.
- David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. ISSN 1476-4687. doi: 10.1038/323533a0. URL <https://doi.org/10.1038/323533a0>.
- Allan Sales, Albin Zehe, Leandro Balby Marinho, Adriano Veloso, Andreas Hotho, and Janna Omeliyanenko. Assessing media bias in cross-linguistic and cross-national populations. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):561–572, 2021. doi: 10.1609/icwsm.v15i1.18084. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/18084>.

- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, October 01, 2019 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv191001108S>.
- Timo Schick and Hinrich Schütze. It’s not just size that matters: Small language models are also few-shot learners, September 01, 2020 2020. URL <https://ui.adsabs.harvard.edu/abs/2020arXiv200907118S>.
- Himanshu Sinha. *A Comparative Review on Sentiment Analysis of Product Review Data Through Large Language Models (LLMs)*. 2024.
- Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. Political ideology and polarization of policy positions: A multi-dimensional approach, June 01, 2021 2021. URL <https://ui.adsabs.harvard.edu/abs/2021arXiv210614387S>.
- E.A. Smith and R.J. Senter. *Automated Readability Index*. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, 1967. URL https://books.google.pt/books?id=vuZD9Q3g2_sC.
- Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: Pretrained bert models for brazilian portuguese. *Intelligent Systems*, pages 403–417. Springer International Publishing, 2020. ISBN 978-3-030-61377-8.
- T. Spinde, L. Rudnitckaia, K. Sinha, F. Hamborg, B. Gipp, and K. Donnay. Mbic – a media bias annotation dataset including annotator characteristics, May 01, 2021 2021a. URL <https://ui.adsabs.harvard.edu/abs/2021arXiv210511910S>.
- Timo Spinde, Felix Hamborg, and Bela Gipp. Media bias in german news articles: A combined approach. *ECML PKDD 2020 Workshops*, pages 581–590. Springer International Publishing, 2020. ISBN 978-3-030-65965-3.
- Timo Spinde, Lada Rudnitckaia, Felix Hamborg, and Bela Gipp. Identification of biased terms in news articles by comparison of outlet-specific word embeddings. *Diversity, Divergence, Dialogue*, pages 215–224. Springer International Publishing, 2021b. ISBN 978-3-030-71305-8.
- 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, 2021c. ISSN 0306-4573. doi: <https://doi.org/10.1016/j.ipm.2021.102505>. URL <https://www.sciencedirect.com/science/article/pii/S0306457321000157>.
- Timo Spinde, Christin Jeggle, Magdalena Haupt, Wolfgang Gaissmaier, and Helge Giese. How do we raise media bias awareness effectively? effects of visualizations to communicate bias. *PLOS ONE*, 17(4):e0266204, 2022a. doi: [10.1371/journal.pone.0266204](https://doi.org/10.1371/journal.pone.0266204). URL <https://doi.org/10.1371/journal.pone.0266204>.
- 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, September 01, 2022 2022b. URL <https://ui.adsabs.harvard.edu/abs/2022arXiv220914557S>.

- Timo Spinde, Smi Hinterreiter, Fabian Haak, Terry Ruas, Helge Giese, Norman Meuschke, and Bela Gipp. The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias, December 01, 2023 2023. URL <https://ui.adsabs.harvard.edu/abs/2023arXiv231216148S>.
- Jón Karl Stefánsson. *Quantitative Measure of Evaluative Labeling in News Reports: Psychology of Communication Bias studied by Content Analysis and Semantic Differential*. Thesis, UiT Norway's Arctic University, 2014.
- Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation, July 01, 2021 2021. URL <https://ui.adsabs.harvard.edu/abs/2021arXiv210702137S>.
- José Luís Sousa Tavares. *Towards an Automated Media Chart: Framing News Articles with Natural Language Processing Techniques*. Thesis, Universidade do Porto, 2023.
- Márcia Raquel Pinto Teixeira. *Índice de Credibilidade de Conteúdos Noticiosos em Língua Portuguesa para Uso em Ambiente Escolar*. Thesis, ISEP, 2021.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, February 01, 2023 2023. URL <https://ui.adsabs.harvard.edu/abs/2023arXiv230213971T>.
- University of Michigan. News bias explored. URL <https://websites.umich.edu/~newsbias/>.
- Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL <http://jmlr.org/papers/v9/vandermaaten08a.html>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, June 01, 2017 2017. URL <https://ui.adsabs.harvard.edu/abs/2017arXiv170603762V>.
- Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp, August 01, 2019 2019. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv190807125W>.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, June 01, 2022 2022a. URL <https://ui.adsabs.harvard.edu/abs/2022arXiv220607682W>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, January 01, 2022 2022b. URL <https://ui.adsabs.harvard.edu/abs/2022arXiv220111903W>.

- Zehao Wen and Rabih Younes. Chatgpt v.s. media bias: A comparative study of gpt-3.5 and fine-tuned language models, March 01, 2024 2024. URL <https://ui.adsabs.harvard.edu/abs/2024arXiv240320158W>.
- Martin Wessel, Tomáš Horych, Terry Ruas, Akiko Aizawa, Bela Gipp, and Timo Spinde. Introducing mbib – the first media bias identification benchmark task and dataset collection, April 01, 2023 2023. URL <https://ui.adsabs.harvard.edu/abs/2023arXiv230413148W>.
- Melvin Wevers. *Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990*. Association for Computational Linguistics, Florence, Italy, 2019. doi: 10.18653/v1/W19-4712. URL <https://aclanthology.org/W19-4712/https://doi.org/10.18653/v1/W19-4712>.
- David Manning White. The “gate keeper”: A case study in the selection of news. *Journalism Quarterly*, 27(4):383–390, 1950. doi: 10.1177/107769905002700403. URL <https://journals.sagepub.com/doi/abs/10.1177/107769905002700403>.
- Alden Williams. Unbiased study of television news bias. *Journal of Communication*, 25(4): 190–199, 2006. ISSN 0021-9916. doi: 10.1111/j.1460-2466.1975.tb00656.x. URL <https://doi.org/10.1111/j.1460-2466.1975.tb00656.x>.
- Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1:270–280, 1989.
- Stephane Wolton. Are biased media bad for democracy? *American Journal of Political Science*, 63(3):548–562, 2019. ISSN 0092-5853. doi: <https://doi.org/10.1111/ajps.12424>. URL <https://doi.org/10.1111/ajps.12424>.
- Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60(2):617–663, 2019. ISSN 0219-3116. doi: 10.1007/s10115-018-1236-4. URL <https://doi.org/10.1007/s10115-018-1236-4>.
- A. Zahid, M. Nasir Khan, A. Latif Khan, F. Kamiran, and B. Nasir. Modeling, quantifying and visualizing media bias on twitter. *IEEE Access*, 8:81812–81821, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2990800.
- Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [mask]: Learning vs. learning to recall, April 01, 2021 2021. URL <https://ui.adsabs.harvard.edu/abs/2021arXiv210405240Z>.
- Xiang Zhou, Heba Elfardy, Christos Christodoulopoulos, Thomas Butler, and Mohit Bansal. *Hidden Biases in Unreliable News Detection Datasets*. Association for Computational Linguistics, Online, 2021. doi: 10.18653/v1/2021.eacl-main.211. URL <https://aclanthology.org/2021.eacl-main.211/https://doi.org/10.18653/v1/2021.eacl-main.211>.
- Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts, June 01, 2023 2023. URL <https://ui.adsabs.harvard.edu/abs/2023arXiv230604528Z>.

Emrullah Şahin, Naciye Nur Arslan, and Durmuş Özdemir. Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. *Neural Computing and Applications*, 2024. ISSN 1433-3058. doi: 10.1007/s00521-024-10437-2. URL <https://doi.org/10.1007/s00521-024-10437-2>.