NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

MORAL FOUNDATIONS THEORY

Morality in Political Discourse in Portugal

Jaime Olívio Teixeira Duarte

Dissertation

presented as partial requirement for obtaining the Master's degree program in Data Science and Advanced Analytics, with specialization in Business Analytics.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Moral Foundations Theory: Morality in Political Discourse in Portugal.

by

Jaime Olivio Teixeira Duarte

Dissertation for obtaining the Master's degree in Data Science and Advanced Analytics, with Specialization in Business Analytics.

Supervised by

Flávio Pinheiro, PhD, NOVA Information Management School
Mafalda Zúquete, MSc, University College of Dublin

July 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the rules of conduct and code of honor from the NOVA Information Management School.

Jaime Olívio Teixeira Duarte

Lisbon, 15 July 2024

DEDICATION

Firstly, I would like to dedicate this dissertation this to my lovely wife Ana Romero and my beautiful children Kamuhelo Duarte, Wami Duarte, Jayana Duarte, and Mayara Duarte. In seeking to provide a better future for them, I had to sacrifice family time when I moved briefly to Portugal. Secondly, I dedicate this work to my dear parents, Agostinho Duarte and Gabriela Teixeira, for their special dedication during my childhood before I left Luanda, Angola, for Windhoek, Namibia, to begin my academic journey. It was through this journey that my desire to continue investing in education was born because knowledge is the best tool any person should seek. To my sister, Nhari Duarte, and my lovely nephew Gabriel Utima Val, I also dedicate this work to you because I always carry you in my heart. Finally, this dissertation is dedicated to all those who have prayed and always believed in me. I will not forget them, as they are part of my challenges and achievements.

ACKNOWLEDGEMENTS

First and foremost, I thank the almighty God who has been my best refuge at all times of my life and in particular during my academic journey. Secondly, I want to thank my supervisor Flavio Pinheiro for accepting me as his supervisee, his patience, understanding and above all, great supervision until it satisfied the desired scientific quality for obtaining a master's degree in data science and advanced Analytics. Thirdly, my other major thanks go to my co-supervisor, Mafalda Zúquete whose, intelligence and expertise in this research area, played a major role in the quality of my work.

To my parents, Agostinho Duarte and Gabriela Teixeira, I wish the Lord our God blesses and keeps them because there is no invoice that pays all their dedications to the person I am today.

To my best friends, Paulo Correia, Fabio Rendel, João Gameiro, Delcio Rodrigues, Osvaldo Amaral, Rogerio Domingos words fail to express the unconditional love we have for each other I wish many blessings from God, along with their wonderful family. To my dear team who stepped up as true workmates, Licinio Mancoca, Valdemira Gomes, Celsio Cosme, Unitel Analytical DIN team, DNC and DSI team´s, my sincere gratitude. To conclude, to all who directly or indirectly made all my journey possible, my sincere gratitude.

ABSTRACT

This study delves into the realm of Moral Foundation Theory (MFT) that are fundamental essential to understand the difference on moral principles between conservative and liberals offering a difference perspective inside US political system; but outside of the America are less clear country with less clear political diversification system that does not follow the same political on conservative and liberals. On this document we will use moral foundation theory and the Portuguese European moral foundation dictionary developed by Mafalda Zuquete to analyze 6 months of articles published by the opinion´s makers from the 2 biggest web newspaper from the country using standard Data Science and Text Mining Techniques to explore its influence on the political options in Portugal on different subjects, emphasizing how moral values shape political narratives and to understand if the follows the same political shape as the one in USA.

KEYWORDS

Moral Foundation Theory; Moral Foundation Dictionary; Political Ideologies in Portugal; Qualitative Political Analysis, Liberals, Conservatives, Text Mining.

Sustainable Development Goals (SGD): https://sdgs.un.org/goals,

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

MFT – Moral Foundation Theory.

JH – Jonathan Haidt.

PSC – Political Science.

PD – Political Discourse.

EC – European Commission.

EU – European Union.

CA – Content Analysis.

PEV – The Green Party (Partido Ecologista "Os Verdes").

SPSS – Statistical Package for the Social Sciences.

ANOVA – Analysis of Variance.

MFD – Moral Foundation Dictionary.

PT – Portugal.

LDA – Latent Dirichlet Allocation.

TF-IDF – Term Frequency-Inverse Document Frequency.

PCA – Principal Component Analysis.

CRISP-DM – Cross-Industry Standard Process for Data Mining.

NLP – Natural Language Processing.

CS – Cosine Similarity

MSD – Moral Scalled Dimensions

KM – Kmeans Cluster Solution

HC – Hierarchical Cluster Solution.

2C – Two Clusters.

3C – Three Clusters.

4C – Four Clusters.

1. INTRODUCTION

Moral Foundations Theory, developed by Jonathan Haidt and Jesse Graham (2007), provides a robust framework for understanding the psychological underpinnings of human morality. The theory comprehends that several innate psychological systems form the basis of our "intuitive ethics," upon which cultures construct diverse moral narratives and institutions. MFT has been particularly influential in examining the moral values and ideological divides between conservatives and liberals in the United States.

This thesis aims to apply the MFT to understand the landscape of political discourse in Portugal, focusing on opinion pieces published by a broad list of authors in mainstream newspapers in Portugal. By analyzing the opinions expressed by political authors over six months, this research seeks to determine the presence and prevalence of different moral foundations in the Portuguese political discourse. It also aims to analyze if the moral dimensions in Portugal align with the patterns observed in the United States, particularly regarding the division between conservative and liberal ideologies.

We will first leverage text-mining techniques to identify the main topics in the corpora of articles being studied, and then we will leverage the European Portuguese Moral Foundations Dictionary developed by Mafalda Zúquete (2023) to identify moral words and how they distribute over the different moral dimensions per author. Evaluating the consistency of moral opinions among Portuguese opinion makers across various topics and examining the authenticity and consistency of the moral positions held by these opinion makers are also key objectives.

Understanding the moral foundations of political discourse in Portugal offers several significant insights. It provides a deeper understanding of the moral landscape in Portuguese society, contributing to the broader knowledge of cultural differences in moral values. It also offers a nuanced view of ideological divides in Portugal, potentially informing political strategies and public policy. Additionally, this research extends the application of Moral Foundations Theory to a new cultural context, enriching the academic discourse on moral psychology and political ideology. Practical implications include helping to evaluate the influence and credibility of opinion makers, affecting public trust and engagement.

2. LITERATURE REVIEW

Moral Foundations Theory (MFT), developed by team of social and cultural psychologists (Haidt & Graham, 2007), seeks to understand why morality, despite cultural differences, often shares common themes and similarities across populations. MFT posits that several innate psychological systems form the basis of our "intuitive ethics." Cultures then construct virtues, narratives, and institutions upon these foundational systems, leading to the diverse moral beliefs observed globally and even within nations. Importantly, MFT provides a descriptive account of human morality rather than a normative one. In this context, evolution does not prioritize the inherent goodness of psychological systems but focuses on creating systems that promote cooperation, survival, and reproduction.

Understanding which side is "morally right" can be challenging in a world where opposing beliefs coexist. The psychologists Haidt and Joseph (2008), proposed a structured approach based on our intuitions and their cultural adaptations to explore the coexistence of divergent moralities, giving rise to what is now known as Moral Foundations Theory. MFT presuppose that human morality is built on several innate pillars that capture society's moral diversity.

Jonathan Haidt's Moral Foundations Theory has been pivotal in understanding moral values, particularly between conservatives and liberals in the United States. Recent studies have applied text mining to analyze political discourse, offering insights into polarization within political views.

The original framework of MFT identified five foundations strongly supported by evidence across various cultures:

1. Care/Harm: This dimension makes us sensitive to signs of suffering and need, fostering a disdain for cruelty and a desire to care for those in distress.
2. Fairness/Cheating: It heightens our awareness of potential collaborators' trustworthiness, driving us to shun or punish cheaters.
3. Loyalty/Betrayal: It attunes us to signals of team allegiance, promoting trust and reward for loyalty while inciting punishment for betrayal.
4. Authority/Subversion: This dimension makes us sensitive to social rank and proper behavior according to status, encouraging respect for authority.
5. Sanctity/Degradation: It encompasses the behavioral immune system, causing us to invest in objects and practices with strong symbolic values, essential for group cohesion.

Initially, the Fairness foundation was skewed toward concerns about equality, which politically left-leaning individuals more strongly endorse in different cultures. In 2011, based on new data, the emphasis shifted towards proportionality, often endorsed by everyone but slightly more by politically right-leaning people. In 2023, based on a decade of empirical work, the

team led by Mohammad Atari (2023) decided to split the Fairness foundation into Equality and Proportionality, making the case for six main foundations:

1. Equality: Defined as "intuitions about equal treatment and equal outcome for individuals."
2. Proportionality: Defined as "intuitions about individuals getting rewarded in proportion to their merit or contribution."

Since MFT was first described by Haidt and Joseph (2004), efforts have been made to identify the candidate foundations for which empirical evidence was strongest. Five criteria for foundation hood (Graham et al., 2013) were proposed:

1. Common in third-party normative judgments.
2. Automatic affective evaluations.
3. Culturally widespread though not necessarily universal.
4. Evidence of innate preparedness.
5. A robust pre-existing evolutionary model.

Several other strong candidates for "foundation hood" have emerged, including:

1. Liberty: This foundation is about feelings of reactance and resentment toward those who dominate and restrict liberty. Its intuitions often conflict with the authority foundation (Iyer et al., 2012), motivating people to oppose oppressors.
2. Honor: This foundation concerns one's self-worth based on reputation and the assessment of others. In Middle Eastern cultures (Atari et al., 2020), honor can lead to expectations of protecting kin and family and retaliating against insults to the family's reputation.
3. Ownership: This foundation has been recognized by moral psychologists for a long time but remains understudied. Ownership intuitions are quick and ubiquitous in human societies, with parallels in other animals. Respect for property is an evolutionarily stable strategy, potentially meeting all criteria for foundation hood (Atari & Haidt, 2023).

MFT was first developed from a review of current evolutionary thinking about morality and cross-cultural research on virtues. It extends the theory of the "three ethics" (Shweder et al., 1997) commonly used globally to discuss morality. The theory is also influenced by relational models theory (Rai & Fiske, 2011).

In summary, MFT provides a comprehensive framework for understanding the psychological underpinnings of human morality. Developed by Jonathan Haidt and Jesse Graham, MFT elucidates how innate psychological systems give rise to the diverse moral beliefs observed across cultures. The theory's evolution, from its initial five foundations to including Equality and Proportionality, underscores its adaptability and relevance in contemporary moral

psychology. Additionally, exploring potential new foundations like Liberty, Honor, and Ownership suggests that MFT continues to expand and refine our understanding of morality.

The influence of MFT extends beyond academic discourse, offering valuable insights into political polarization and the moral dimensions of societal issues. MFT bridges the gap between universal human traits and cultural variations by grounding moral values in evolutionary and cross-cultural perspectives. This theory enhances our understanding of moral diversity and provides a robust framework for analyzing moral behavior in various contexts.

2.1. RELATED WORK

In recent years, several studies have applied text-mining and data-mining techniques to analyze political discourse through the lens of MFT, providing insights into the moral dimensions of political communication across different countries and languages.

For instance, a group of Portuguese researchers Mafalda Zúquete, Flávio L. Pinheiro, and Diana Orghian (2023) developed a Portuguese version of the Moral Foundations Dictionary (MFD) and applied it to analyze parliamentary debates in Portugal. The researchers utilized text mining techniques to identify and categorize moral language, offering insights into the moral rhetoric of Portuguese politicians. They employed natural language processing (NLP) to align sentences with the Portuguese MFD, thus quantifying the presence of moral foundations in political speeches. The analysis revealed distinct patterns of moral language usage between political parties, with conservative parties emphasizing binding foundations (authority, loyalty, purity) more than liberal parties, which focused more on individualizing foundations (care, fairness).

Similarly, a study conducted by Hiroki Takikawa and Takuto Sakamoto (2020) explored the differences in moral rhetoric between the U.S. and Japan by analyzing legislative speeches using MFT. Text mining and sentiment analysis were employed to process and analyze the speeches, extracting moral language patterns and correlating them with emotional tones. The study found that U.S. politicians frequently invoked individualizing foundations, while Japanese politicians emphasized binding foundations, reflecting broader cultural differences in moral priorities between the two countries.

The extended Moral Foundations Dictionary (eMFD) project, involving researchers such as Frederic Hopp, Fisher Jacob, Cornell Devin , Huskey Richard and Rene Webber (2020), aimed to enhance the accuracy of moral content analysis by incorporating crowd-sourced annotations and expanding the dictionary to cover more languages and cultural contexts. Data-driven techniques and machine learning were utilized to validate and refine the dictionary entries. The eMFD has been applied to texts in multiple languages, demonstrating its flexibility and effectiveness in capturing moral intuitions from large text corpora.

In another social media study, J. Hoover, M. Dehghani, and colleagues (2020) analyzed moral language on Twitter using text mining techniques. They annotated tweets for moral sentiment and applied the MFD to classify them according to different moral foundations. The analysis revealed that moral language on Twitter often reflects immediate and intuitive moral reactions, with significant variations depending on users' political orientation.

These recent studies illustrate the powerful combination of MFT, data science, and text mining techniques in uncovering the moral underpinnings of political discourse across different languages and cultural contexts. Researchers can reveal the complex moral landscapes that shape political communication by developing and applying sophisticated text analysis tools. These methodologies and findings provide a valuable foundation for analyzing the morality of Portuguese parties' political discourse, offering theoretical insights and practical techniques to guide our research. Integrating MFT into this research will help elucidate the moral foundations underlying political ideologies and contribute to a deeper understanding of the dynamics driving moral and ethical judgments in contemporary society.

3. DATA AND METHODS

3.1. DATA

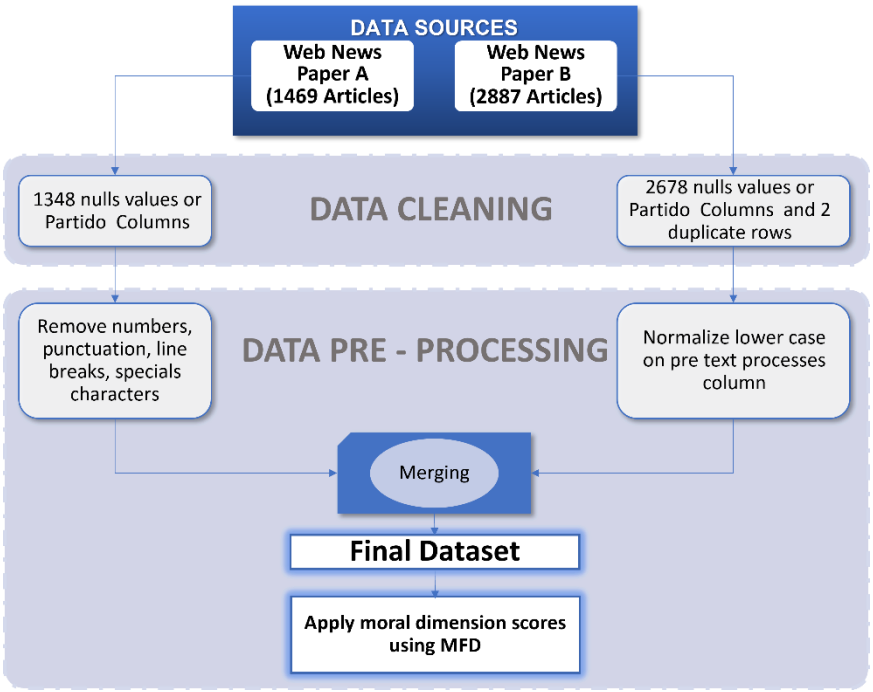

Figure 3.1. Data collection process

In this study, we conducted a manual approach to collect data from two of Portugal's most relevant mainstream newspapers. The collected data was meticulously cleaned, processed, transformed, and encoded separately before being merged into a final dataset. Subsequently, we employed several advanced techniques to analyze the data: tokenizing and lemmatizing the news content, applying topic modelling to identify the most relevant topics, manually classifying the words using a moral foundation dictionary to quantify each moral dimension, and utilizing unsupervised learning clustering algorithms to profile the authors. These two datasets can be used in future for other types of research purposes considering the period that was extracted.

To investigate political discourse morality in Portugal, we manually extracted articles published by opinion makers over six months, from October 1, 2023, to March 31, 2024, from two of the most prominent national web newspapers. This effort culminated in the creation of two structured datasets, one for each source, that correspond to:

- Newspaper A comprised 1,469 articles with detailed metadata, including the publication date, time, political party mentioned, title, content, link, author, subtitle, type, editor, number of shares, and comments.
- Newspaper B contained 2,887 articles with a similar structure, though it notably lacked values in the 'Partido' (political party) column.

After assembling these datasets, we conducted a quality analysis to identify duplicates and null values. Newspaper A had one duplicate row and significant missing values in columns such as 'Partido' and 'Sub-Titulo' (Subtitle). Newspaper B had two duplicates and numerous missing values in similar columns. Given the high rate of missing values, we decided to drop the 'Partido' and 'Sub-Titulo' columns from both datasets. This decision was driven by the need to maintain data integrity and ensure the remaining data was as complete and useful as possible.

We cleaned the data by removing special characters from the 'Titulo' (Title) and 'Author' columns. This step was crucial to standardize the textual data and facilitate accurate analysis. Subsequently, we standardized the formats of all columns, converting 'Data' to dates, 'Author' to strings, and 'Partilhas_QTD' (Sharing Quantity) and 'Comentarios_QTD' (Comments Quantity) to integers, ensuring the datasets were ready for subsequent steps. This preprocessing ensured all variables were in appropriate formats, essential for reliable data analysis.

Next, we engineered new features to enhance the datasets. We created a 'Week' column from the 'Data' column to capture the week number and a 'Month' column to indicate the month. Additionally, we added a 'Month_Year' column to specify the year and a 'Month_Name' column to map the month number to its name. we also created a 'Total_Words' column to count the number of words in each article. These new features provided additional temporal context and facilitated detailed time-series analysis.

These enhanced datasets were then combined into a final dataset with columns: 'Data', 'Titulo', 'Noticia' (Article), 'Link', 'Author', 'Month_name', 'Week', 'Editor', 'Total_Words', 'Month_Year'. This consolidation aimed to streamline the analysis by providing a comprehensive view of the collected data. The final combined dataset revealed that articles were distributed 64% from Newspaper B and 36% from Newspaper A.

To prepare for moral foundation analysis, we created a 'PreProcessedText' column from the 'Noticia' column. This column underwent extensive cleaning to remove HTML tags, punctuation, numbers, line breaks, and special characters. The text was normalized to lowercase, and multiple spaces were reduced to single. This pre-processing was critical to ensure that the textual data was uniform and noise-free, thereby enhancing the accuracy of subsequent text analysis.

Using Mafalda Zuquete (2023) Portuguese European moral foundation dictionary, a comprehensive lexical resource that maps words to various moral dimensions, we calculated moral dimension scores for each article. This dictionary is meticulously curated to reflect words associated with the five moral foundations divided into virtue and vice categories: Harm/Care, Fairness/Reciprocity, Ingroup/Loyalty, Authority/Respect, and Purity/Sanctity. Leveraging this dictionary, we generated eight new columns in the dataset: 'HarmVirtue', 'HarmVice', 'FairnessVirtue', 'FairnessVice', 'IngroupVirtue', 'IngroupVice', 'AuthorityVirtue', 'AuthorityVice', 'PurityVirtue', and 'PurityVice'. We decide to not consider the

“Ingroup/Loyalty” ('IngroupVice'/'IngroupVirtue') for this research. These columns represent the counts of words related to each specific moral foundation dimension within the articles.

Additionally, we computed 'Total_Classified_words' to correspond the number of words from each article identified and classified according to the moral foundations. The 'Classified' column was added to indicate whether an article contained any words that could be classified into the moral foundations. This comprehensive classification was pivotal in quantifying the moral content of the articles, enabling a detailed and structured analysis of moral foundations in political discourse. By quantifying these moral dimensions, we could systematically investigate how different moral values are emphasized or downplayed in the political discourse of Portuguese opinion-makers.

These calculated scores provided critical insights into the moral landscape of the political articles, setting a strong foundation for further detailed analysis and interpretation in subsequent sections of this study.

We also end up using, the Pareto Principle, also known as the 80/20 rule, was employed to enhance the focus on the most prolific authors. By filtering authors with a minimum threshold of two articles, the dataset was refined, resulting in a substantial reduction in the number of unique authors from 1162 to 356 while retaining approximately 80.94% of the total articles. This strategic filtering demonstrated that a smaller subset of authors contributes to the majority of the content, thereby optimizing the dataset for further analysis.

Table 3.1. Data Schema Overview

Column Name	Description	Nature of Information	Origin
Author	Author of the article	Categorical (Text)	Original
Comentarios_QTD	Comments quantity	Numeric (Discrete)	Original
Data	Date of the article	Categorical (Date)	Original
Editor	Editor of the article	Categorical (Text)	Original
Link	URL of the article	Text	Original
Noticia	Content of the article	Text	Original
Partilhas_QTD	Sharing quantity	Numeric (Discrete)	Original
Título	Title of the article	Text	Original
Week	Week of publication	Categorical (Numeric)	Original
AuthorityVice	Number of words classified as Authority Vice	Numeric (Discrete)	Added

Column Name	Description	Nature of Information	Origin
AuthorityVirtue	Number of words classified as Authority Virtue	Numeric (Discrete)	Added
Classified	Indicator if the article is classified	Categorical (Binary)	Added
FairnessVice	Number of words classified as Fairness Vice	Numeric (Discrete)	Added
FairnessVirtue	Number of words classified as Fairness Virtue	Numeric (Discrete)	Added
HarmVice	Number of words classified as Harm Vice	Numeric (Discrete)	Added
HarmVirtue	Number of words classified as Harm Virtue	Numeric (Discrete)	Added
IngroupVice	Number of words classified as Ingroup Vice	Numeric (Discrete)	Added
IngroupVirtue	Number of words classified as Ingroup Virtue	Numeric (Discrete)	Added
Month_name	Name of the month of publication	Categorical (Text)	Added
Month_Year	Month and year of publication	Categorical (Date)	Added
MoralityGeneral	Number of words classified as Morality General	Numeric (Discrete)	Added
PreProcessedText	Preprocessed content of the article	Text	Added
PurityVice	Number of words classified as Purity Vice	Numeric (Discrete)	Added
PurityVirtue	Number of words classified as Purity Virtue	Numeric (Discrete)	Added
Titulo_org	Original title of the article	Text	Added
Total_Classified_Words	Total number of classified words in the article	Numeric (Continuous)	Added
Total_Words	Total number of words in the article	Numeric (Continuous)	Added

The final dataset comprised 3,297 articles and 2,378,103 words, with 72,358 words classified according to moral dimensions. The authorship was diverse, with 356 unique authors contributing to the discourse. The distribution of articles among the top 30 authors revealed that a significant portion of the content was concentrated among a few prolific authors. For instance, Miguel Esteves Cardoso contributed 6.2% of the articles, followed by Henrique Raposo with 3.2%. The top 30 authors accounted for 41% of the total articles, highlighting their dominant role in shaping public opinion.

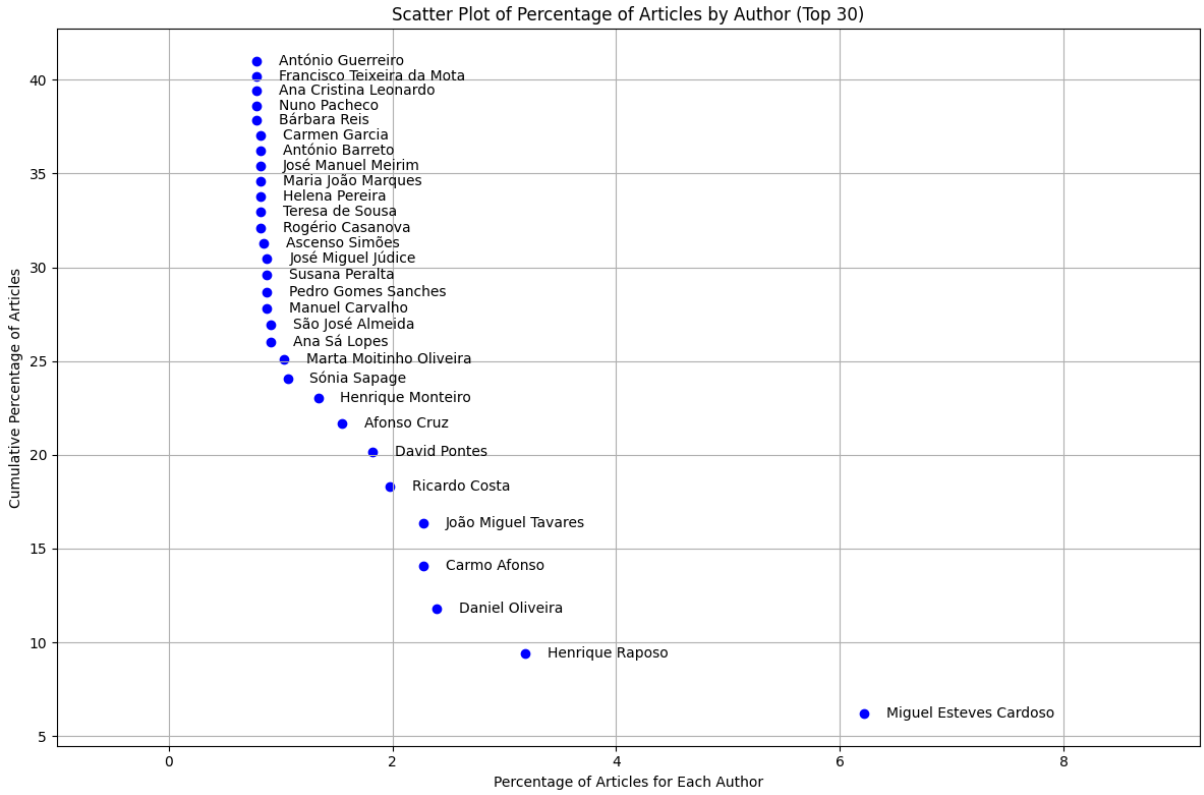

Figure 3.2. Percentage of articles by Author (Top 30)

In terms of word count, the same pattern emerged. Miguel Esteves Cardoso was the most verbose author, contributing 3.5% of the total words, reflecting his extensive engagement in political commentary.

Analyzing the temporal distribution, there was a noticeable peak in December, with a decline towards March. This pattern was consistent in both the number of articles and the total word count but on the total classified words per month indicated a steady stream of morally charged content. October had the highest classified words, possibly due to the heightened political activity following the summer recess.

On weekly breakdown showed fluctuations, with significant spikes corresponding to major political events or crises, highlighting the reactive nature of political journalism (Figure 3.3), and on 2023, W53 as Christmas week, in 2024, W13 as easter season break with the lowest publish articles.

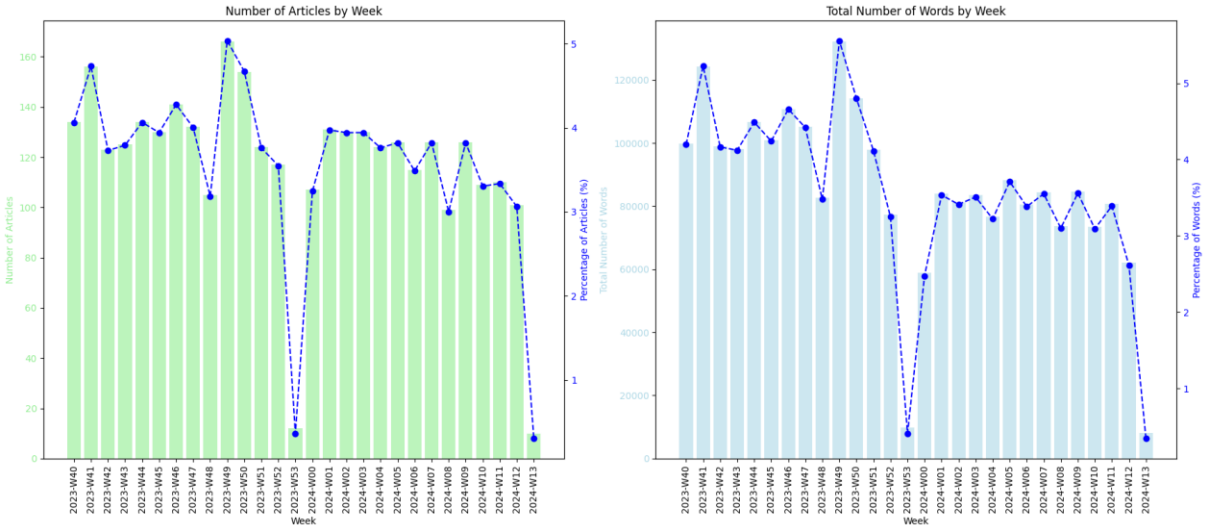

Figure 3.3. Total Number of Articles and Words by Week

The monthly analysis of the top authors revealed dynamic changes in authorship prominence per number of articles published monthly. For instance, Miguel Esteves Cardoso consistently led in article count, but other authors like Carmo Afonso and João Miguel Tavares showed increased activity in certain months, reflecting their engagement with specific political topics (Table 3.3).

Table 3.2. Monthly Top 5 Authors per articles and words percentage

Month	Author	Percentage of Articles	Percentage of Words
October	Miguel Esteves Cardoso	6.4%	13.2%
	Henrique Raposo	3.6%	12.4%
	Daniel Oliveira	2.7%	12.1%
	João Miguel Tavares	2.2%	10.6%
	Carmo Afonso	2.1%	9.3%
November	Miguel Esteves Cardoso	5.1%	13.0%
	Henrique Raposo	3.2%	12.8%
	Daniel Oliveira	2.7%	12.1%
	João Miguel Tavares	2.5%	10.2%
	Carmo Afonso	2.1%	9.6%
December	Miguel Esteves Cardoso	5.1%	13.2%
	Henrique Raposo	2.8%	12.9%
	Daniel Oliveira	2.7%	11.8%

Month	Author	Percentage of Articles	Percentage of Words
	João Miguel Tavares	2.7%	11.6%
	Carmo Afonso	2.6%	10.7%
January	Miguel Esteves Cardoso	6.1%	14.0%
	Henrique Raposo	3.6%	12.6%
	Daniel Oliveira	3.0%	11.2%
	João Miguel Tavares	2.3%	10.5%
	Carmo Afonso	2.3%	9.8%
February	Miguel Esteves Cardoso	6.4%	14.0%
	Henrique Raposo	3.0%	12.8%
	Daniel Oliveira	2.8%	11.0%
	João Miguel Tavares	2.6%	10.6%
	Carmo Afonso	2.5%	9.7%
March	Miguel Esteves Cardoso	7.1%	15.0%
	Henrique Raposo	3.5%	12.8%
	Daniel Oliveira	2.9%	11.2%
	João Miguel Tavares	2.8%	10.6%
	Carmo Afonso	2.5%	10.1%

The monthly evolution of moral foundations dimensions showed that AuthorityVirtue consistently had the highest count, followed by HarmVirtue. The Vice dimensions exhibited a declining trend, particularly HarmVice, suggesting a decrease in negative moral judgments over time.

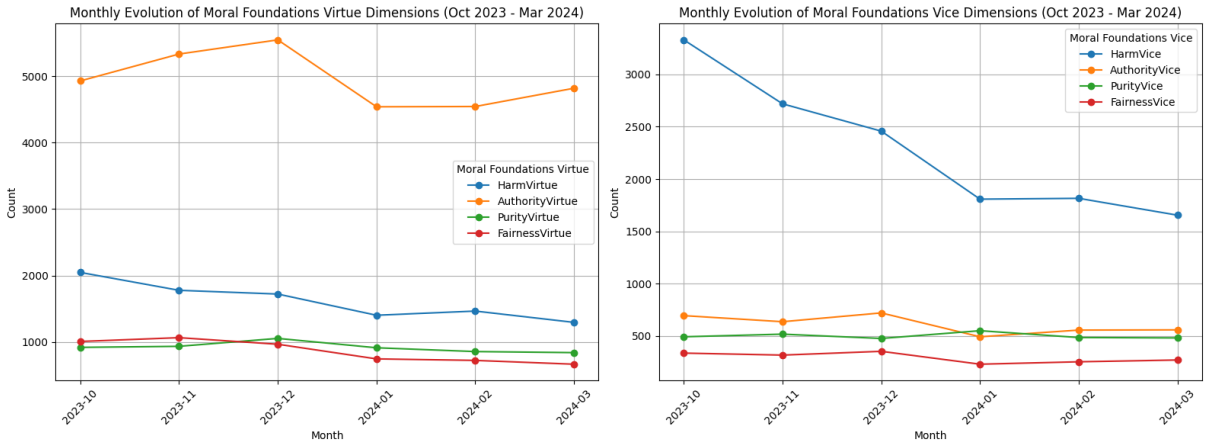

Figure 3.4. Monthly Evolution of Moral Foundations

Descriptive statistics of the final dataset further underscored the variability and richness of the data. The mean number of total words per article was 727.91, with a wide range from 10 to 3428 words, reflecting the diverse lengths of opinion pieces. The moral foundation scores varied significantly, with AuthorityVirtue and HarmVirtue being the most prevalent virtues, while FairnessVice and PurityVice were less frequently mentioned. These statistics provide a comprehensive overview of the dataset, setting the stage for more nuanced analysis.

Table 3.3. Descriptive statistical table

	mean	std	min	25%	50%	75%	max	var
Harm Virtue	3.04	3.88	0.0	1.0	2.0	4.0	88.0	15.08
Harm Vice	4.06	4.58	0.0	1.0	3.0	5.0	44.0	21.00
Authority Virtue	9.08	7.29	0.0	4.0	8.0	13.0	59.0	53.19
Purity Virtue	1.59	2.29	0.0	0.0	1.0	2.0	36.0	5.26
Authority Vice	1.10	1.57	0.0	0.0	1.0	2.0	29.0	2.49
Purity Vice	0.98	1.95	0.0	0.0	0.0	1.0	27.0	3.83
Fairness Virtue	1.62	2.14	0.0	0.0	1.0	2.0	28.0	4.61
Fairness Vice	0.55	1.04	0.0	0.0	0.0	1.0	15.0	1.08
Total Words	727.90	378.40	10.0	481.0	694.0	935.5	3428.0	143190.95
Total Classified Words	22.04	14.95	0.0	11.0	20.0	30.0	140.0	223.63
Classified	0.98	0.11	0.0	1.0	1.0	1.0	1.0	0.01

To conclude the data exploration chapter, we employed the cosine similarity technique to delve deeper into the relationships between articles. By transforming the textual data into numerical vectors using the TF-IDF vectorizer, we could measure the content-based similarity between articles (Rahutomo et al., 2012).

The cosine similarity matrix revealed fascinating insights. We observed a wide range of similarity values, indicating a diverse spectrum of content. While many articles showed low similarity, suggesting a broad variety of topics, there was also a notable cluster of articles with high similarity scores. These high similarity pairs pointed to significant content overlap, hinting at possible duplicate content or closely related themes.

The distribution of similarity values followed a roughly normal pattern, peaking around the 0.6-0.7 range. This indicated that most articles had a moderate level of similarity. The threshold of 0.8, used to identify highly similar articles, was validated by the histogram, which showed a clear drop in frequency beyond this point.

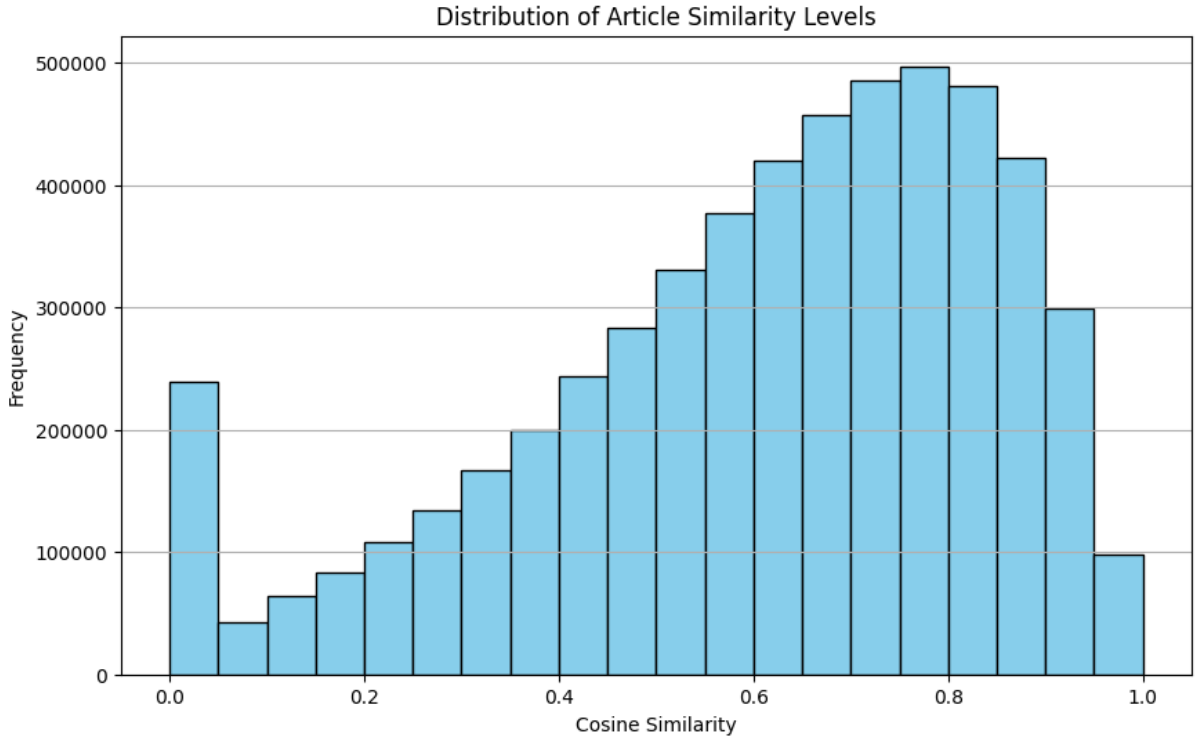

Figure 3.5. Histogram of Cosine Similarity Between Articles

Additionally, the histogram's peaks and troughs suggested potential clusters of related articles, which could be further explored to uncover common themes. The presence of high similarity pairs also highlighted some level of content redundancy, presenting an opportunity to refine the dataset by reducing duplicate information.

This analysis was crucial for laying the groundwork for clustering analysis, enabling us to identify distinct clusters of articles based on their moral content and textual similarity.

3.2. METHODS

This section explores the methodologies and models employed to analyze our dataset, offering insights into the moral underpinnings of political commentary in Portugal. We aim to uncover the intricate patterns and narratives that shape public discourse by delving into the techniques used. A multifaceted methodology was adopted to ensure a comprehensive analysis of the research on the morality of the Portuguese political context. This approach encompasses various methods and techniques, each selected for its suitability to address specific aspects of the research. Below, the methodologies employed throughout the study are outlined and elaborated, providing references and additional context.

To comprehensively analyze the morality in Portuguese political discourse, we adopted a multifaceted methodology. The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was applied to structure the research. CRISP-DM provides a robust and structured framework for data mining projects, ensuring a systematic approach from the initial understanding of the business problem to the deployment of the final model. This methodology was chosen for its wide recognition in data science, offering flexibility and comprehensiveness suitable for complex projects (Durango Vanegas et al., 2023).

The primary goal of the business understanding phase was to comprehend the research objectives and requirements from a business perspective, translating these into a data mining problem. The key objective was to analyze the moral dimensions within the political discourse of Portuguese opinion makers and understand how these dimensions are distributed between conservatives and liberals.

In the data understanding phase, the focus was on collecting initial data and familiarizing with the dataset. The data was gathered from the two biggest web newspaper websites in Portugal using a manual data scraping approach. This data spanned from October 2023 to March 2024 and included various attributes such as date, title, news content, author, and editor. Exploratory data analysis was conducted to understand the structure and distribution of the data, identifying any quality issues.

The data preparation phase involved multiple steps to clean and preprocess the data for modelling. This included handling missing values, creating new variables (e.g., month and

week), and combining datasets into a single comprehensive dataset. Text preprocessing steps such as tokenization, lemmatization, and stop-word removal were applied to the news content. Additionally, the moral foundation scoring methodology was employed to analyze moral foundations within the political discourse. This involved utilizing a Portuguese dictionary developed by Mafalda Zúquete (2023), tailored to capture the nuances of moral dimensions in Portuguese texts. The scoring process involved analyzing texts to count the words associated with each moral dimension, such as Harm, Authority, Purity, and Fairness. This method allows for a quantitative assessment of the prevalence of different moral values in the texts of opinion makers. MFT by Haidt and Graham (2007) underpins this approach, providing a theoretical basis for understanding the role of different moral dimensions in shaping political and social attitudes.

During this phase, the cosine similarity technique was employed using the TF-IDF vectorizer to transform textual data into numerical vectors, enabling the measurement of similarity between articles. This analysis provided insights into the relationships between articles and helped identify clusters of related content. Furthermore, the Pareto principle was applied to filter the data, ensuring that the most significant contributors were retained for analysis.

Before the modelling phase, the Min-Max scaling technique was employed for data scaling to normalize the moral dimensions (Harm, Authority, Purity, and Fairness). Min-Max scaling was chosen because it allows for setting specific scaling values, ensuring that the data is brought onto a common scale without distorting the differences in the values (Vafaei et al., 2020).

The Pearson correlation method was utilized to examine the relationships between different types of variables. Pearson correlation was selected due to its ability to measure the linear relationship between variables, providing insights into the strength and direction of these relationships within our dataset (Hauke & Kossowski, 2011).

In the modelling phase, various techniques were employed to analyze and cluster the articles. Initially, the LDA (Latent Dirichlet Allocation) model was used for topic modelling to identify main topics within the dataset (Lee & Ostwald, 2024). The primary goal was to extract dominant topics from the articles using the LDA. The robustness of the topic model was ensured by evaluating it using coherence scores and perplexity measures, which helped in confirming the quality and interpretability of the topics generated.

Following this, unsupervised clustering was performed using K-means and hierarchical clustering algorithms (Gao et al., 2022). Two approaches were tested on both algorithms: one using cosine similarity on preprocessed article content and the other using scaled moral score columns.

The elbow method was used to identify the ideal number of clusters for K-means, and the dendrogram was utilized for hierarchical clustering to determine the optimal number of clusters. The elbow method helps in pinpointing the point where adding more clusters does not significantly improve the model, while the dendrogram provides a visual representation of the hierarchical clustering process, assisting in identifying the appropriate number of clusters (Shi et al., 2021).

The K-means clustering algorithm (Gao et al., 2022) was employed using cosine similarity on the columns "PreProcessedText" to transform the words into vectors and also on the moral foundation scaled scores to work on the scores on the moral dimensions from the MFT dictionary. This facilitated the visualization of clusters and the creation of an Authors Matrix, mapping authors to their respective clusters. By analyzing the cluster profiles, a deeper understanding of the differences and characteristics of each cluster was gained. K-means is a widely used clustering algorithm due to its simplicity and efficiency, while PCA effectively reduces the dimensionality of data, making it easier to visualize and interpret.

Additionally, hierarchical clustering based on Ward's distance was utilized on the scaled scores variables. Combined with the Authors Matrix and cluster profiles, this method provided a hierarchical perspective on the relationships between authors and the clustering structure, further enriching the analysis (Vichi et al., 2022). Hierarchical clustering is useful for understanding the nested structure of data and identifying sub-clusters within larger groups.

To evaluate the quality of the clustering solutions, the silhouette score (Shutaywi & Kachouie, 2021) and the Davies-Bouldin index (Ros et al., 2023) were employed. The silhouette score measures how similar an object is to its own cluster compared to other clusters, providing an intuitive measure of cluster cohesion and separation. The Davies-Bouldin index evaluates the average similarity ratio of each cluster with the cluster that is most similar to it, ensuring a balance between intra-cluster scatter and inter-cluster separation. These metrics were chosen

to ensure a comprehensive and robust evaluation of the clustering solutions (K-means and hierarchical), facilitating the identification of the most appropriate algorithm for the data.

The deployment phase involved summarizing the findings and preparing the results for presentation. Visualization techniques such as PCA (Principal Component Analysis), percentage matrix, and radar charts were used to represent cluster profiles effectively, aiding in the clear communication of results.

This structured approach ensured a thorough, reproducible, and grounded analysis of morality in Portuguese political discourse. By integrating these methodologies, the research achieved a comprehensive understanding of the moral dimensions in Portuguese political discourse, offering valuable insights into the moral underpinnings of opinion makers in prominent web newspapers. The combination of quantitative and qualitative techniques ensured a thorough exploration of the data, while the structured approach provided a clear framework for analysis and interpretation.

4. TOPIC MODELLING RESULTS

In order to uncover the hidden themes within the rich landscape of Portuguese political discourse, we continued our exploration by creating a new variable “Topic_Name” from the variable “PreprocessedText” using the Latent Dirichlet Allocation (LDA) algorithm for topic modelling (Lee & Ostwald, 2024).

LDA is a powerful technique in text mining that helps identify abstract topics within a collection of documents. This method allows us to discern patterns and thematic structures within large text corpora, providing deeper insights into the underlying narratives and themes present in the political discourse. That´s why starting with the critical task of determining the optimal number of topics, denoted as K, was necessary for our LDA model. To that end, we meticulously computed coherence values, which guide me towards the most interpretable and consistent topics. The coherence plot (Figure 4.1) vividly illustrates how the model's coherence score fluctuates with varying numbers of topics. Although the highest coherence score was achieved with 13 topics, we ultimately chose eight topics for their greater relevance and interpretability in the context of our study.

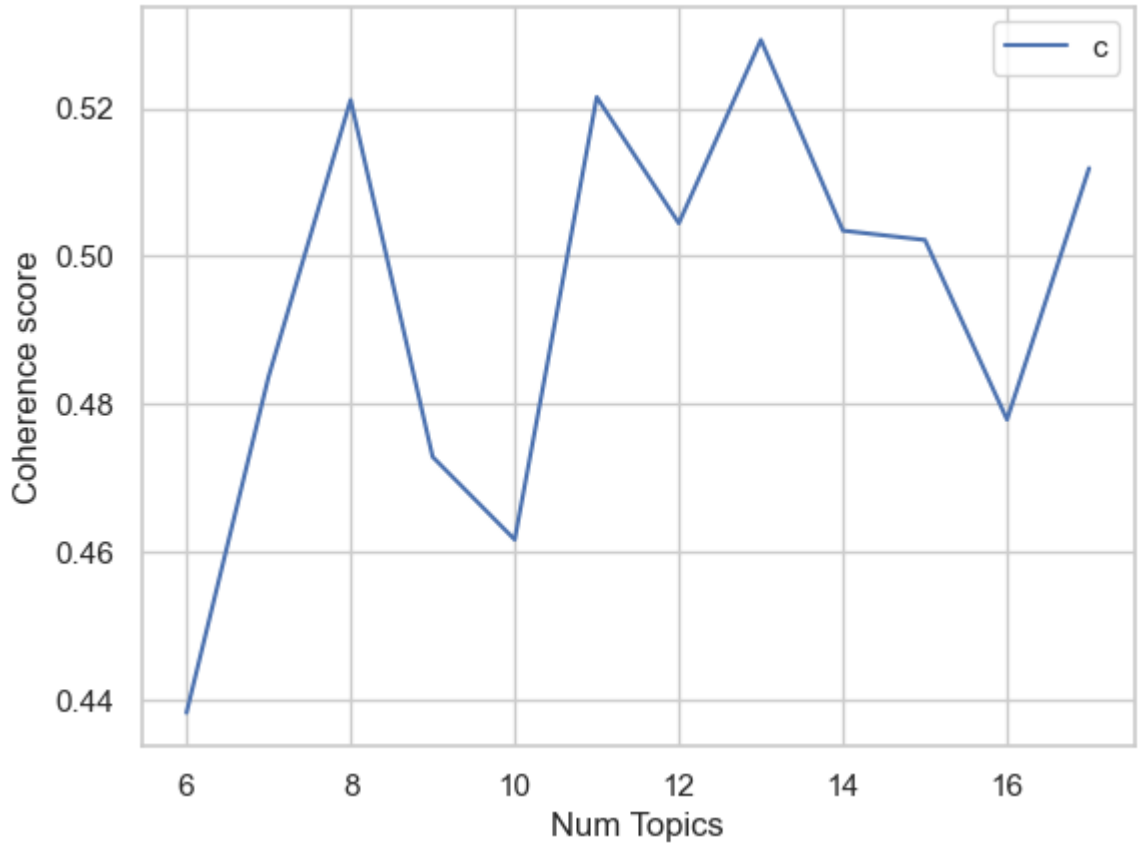

Figure 4.1. Coherence plot to determine the optimal number of topics

With the number of topics set, the LDA model was run on the corpus of opinion articles. The result was a collection of word clouds (Figure 4.1) that visually encapsulated the essence of

each topic, providing an immediate and clear thematic breakdown. As the model processed the articles, eight distinct topics emerged, each characterized by a unique set of keywords. These topics paint a vivid picture of the central themes:

1. **Topic 1:** Words like "fazer", "todo", "poder", and "dizer" dominate, hinting at general discussions.
2. **Topic 2:** With words like "filmar", "desviar", and "herói", this topic touches on heroism and notable events.
3. **Topic 3:** This topic, featuring "Israel", "guerra", and "hama", delves into the conflict between Israel and Palestine.
4. **Topic 4:** Centered around "rendimento", "euro", and "fiscal", this topic explores fiscal policies and economic discussions.
5. **Topic 5:** Keywords like "político", "partido", and "ps" focus on national politics and party dynamics.
6. **Topic 6:** Words such as "vinho", "energia", and "energético" highlight discussions on energy and climate.
7. **Topic 7:** Featuring "trocar", "sínodo", and "costumar", this topic covers various general topics and shifts.
8. **Topic 8:** With terms like "poder", "público", and "político", this topic deals with governance and public matters.

The distribution of word counts by dominant topics (Figure 4.2) provides further insights into how these themes are spread across the corpus.

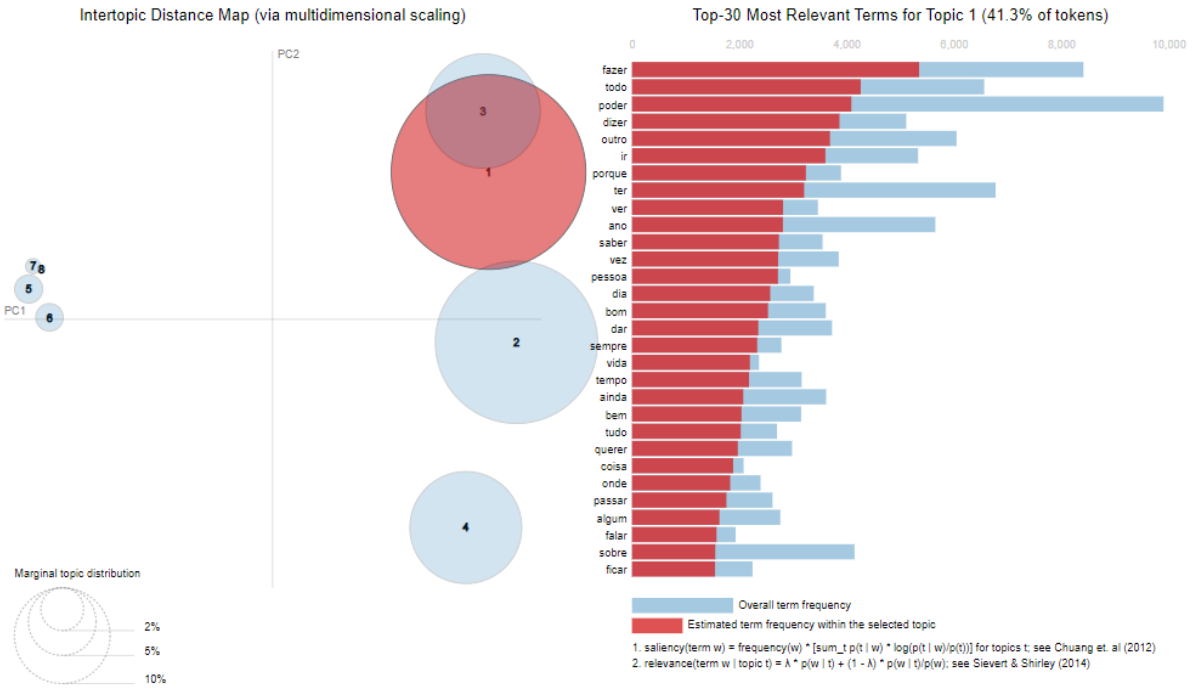

Figure 4.2. Distribution Map for Topic and Most relevant terms for Topic 1

To research deeply, we used PyLDAvis (Islam, 2019) to visualize the contribution of each topic. The analysis revealed the following distribution:

Table 4.1. Tokens percentage of Type of Topics

Nº	Type	Percentage of tokens
Topic 1	General Discussions	41.3%
Topic 2	Reports and Social Issues	28.7%
Topic 3	Israel-Palestine Conflict	14.3%
Topic 4	Economy and Finance	13.7%
Topic 5	National Politics and Parties	0.9%
Topic 6	Climate and Energy	0.8%
Topic 7	Events and Personal Exchanges	0.3%
Topic 8	General Politics and Social Issues	0.1%

The model's assessment metrics attest to its robust performance with a perplexity score of -8.83 and a coherence score of 0.52051 (Tijare & Rani, 2020).

To conclude the LDA model results, we present a summary table highlighting the main topics identified within the dataset. Each topic is characterized by its dominant words and the percentage of the total content it represents. Here is a detailed explanation of the table:

Table 4.2. Percentage of Status about the Topics

Topics	Words	Percentage
General Discussions	Dominating	47.01%
General Politics and Social Issues	Following	29.72%
National Politics and Parties	Representing	13.04%
Israel-Palestine Conflict	Accounting	10.19%
Climate and Energy	Contributing	0.03%

1. **General Discussions:**

This topic accounts for 47.01% of the content, making it the most dominant theme in the dataset. It encompasses broad discussions that may not be specific to any particular issue but are important for setting the context of political and social discourse.

2. **General Politics and Social Issues:**

Representing 29.72% of the content, this topic covers a wide range of political and social matters. It includes debates and discussions on various issues affecting society and general political discourse.

3. **National Politics and Parties:**

This topic, making up 13.04% of the content, focuses specifically on national politics, including discussions about political parties and their activities. It highlights the dynamics and narratives within the national political landscape.

4. **Israel-Palestine Conflict:**

Accounting for 10.19% of the content, this topic is centered on the Israel-Palestine conflict. It includes commentary, analysis, and reporting related to this specific geopolitical issue.

5. **Climate and Energy:**

Although it only represents 0.03% of the content, this topic addresses discussions around climate change and energy policies. Despite its small percentage, it is a crucial area of discourse with significant implications for policy and public opinion.

By analyzing these topics, the LDA model has provided a structured understanding of the major themes present in the dataset, allowing for deeper insights into the moral and political dimensions of the articles. This comprehensive topic analysis sets the stage for further exploration and clustering based on the identified themes.

5. CLUSTERING SOLUTION RESULTS

In this section, we present a comparative analysis of four clustering solutions applied to understand the moral dimensions in political discourse in Portugal, aiming to identify the best approach among them. The approaches examined include K-means with Cosine Similarity, K-means with Moral Dimensions, Hierarchical Clustering with Cosine Similarity, and Hierarchical Clustering with Moral Dimensions. Each method was evaluated to identify clear distinctions among clusters and determine the optimal solution.

Each clustering technique and metric was assessed using various evaluation metrics to ensure robustness and validity. The Elbow Method and dendrograms were used to determine the optimal number of clusters. The Silhouette Score and Davies-Bouldin Index provided measures of cluster cohesion and separation, indicating the effectiveness of each clustering approach. The results are summarized in Table 5.1 below:

Table 5.1. Comparative Analysis of Clustering Techniques for Moral Dimensions in Portuguese Political Discourse

Metric	Cluster Metric		Cluster Evaluation	
	Elbow Method	Dendrogram	Silhouette Score	David Watson
Cosine	2		0.4101	1.2214
	3		0.2790	1.3513
	4		0.2730	1.3472
Euclidean	2		0.3359	1.5036
	3		0.3327	1.5446
Cosine		2	0.1488	0.8595
		3	0.1450	0.8606
Euclidean		2	0.3454	1.6471
		3	0.3427	1.8093

Based on the results, it is evident that the K-means clustering approach using Cosine Similarity and the Hierarchical Clustering approach using Euclidean distance yielded the best results. However, among these, the Hierarchical Clustering with Euclidean distance provided more meaningful cluster distributions. The evaluation metrics, including the Silhouette Score and Davies-Bouldin Index, support this conclusion, indicating robust cluster cohesion and separation.

Finally, the two-cluster solution obtained through the hierarchical clustering approach using Euclidean distance emerged as the optimal solution. This method provided the most robust framework for understanding the separation among authors and the moral dimensions in the opinion articles. To be more detailed on the results for this approach, before running the cluster solution we ensured the moral dimensions were properly scaled and exhibited lower

positive correlations. This step was crucial to enhance the uniqueness and independence of each moral dimension, reducing redundancy and multicollinearity. To enhance the analysis, we focused on the dimensions of 'HarmVirtue,' 'AuthorityVirtue,' 'PurityVirtue,' 'FairnessVirtue,' 'HarmVice,' 'AuthorityVice,' 'PurityVice,' and 'FairnessVice.' Scaling these dimensions was essential to ensure consistency and equal weighting in our analysis. We generated a correlation matrix to understand the relationships between these dimensions, which was a critical preparatory step before applying any clustering algorithm.

Figure 5.1. Spearman Correlation Matrix of Scaled Moral Dimensions

The Spearman better correlation matrix (Figure 5.1) revealed no excessively high correlations between the dimensions. Each dimension provided unique information, which is crucial for the accuracy and effectiveness of the clustering process. Lower correlations imply that each moral dimension contributes to the overall analysis, reducing the risk of overlapping information.

Moving forward, the dendrogram for hierarchical clustering offers a visual representation of the clustering process, highlighting how individual points or clusters are merged step by step. This provides an intuitive understanding of the data structure and supports using 2 clusters.

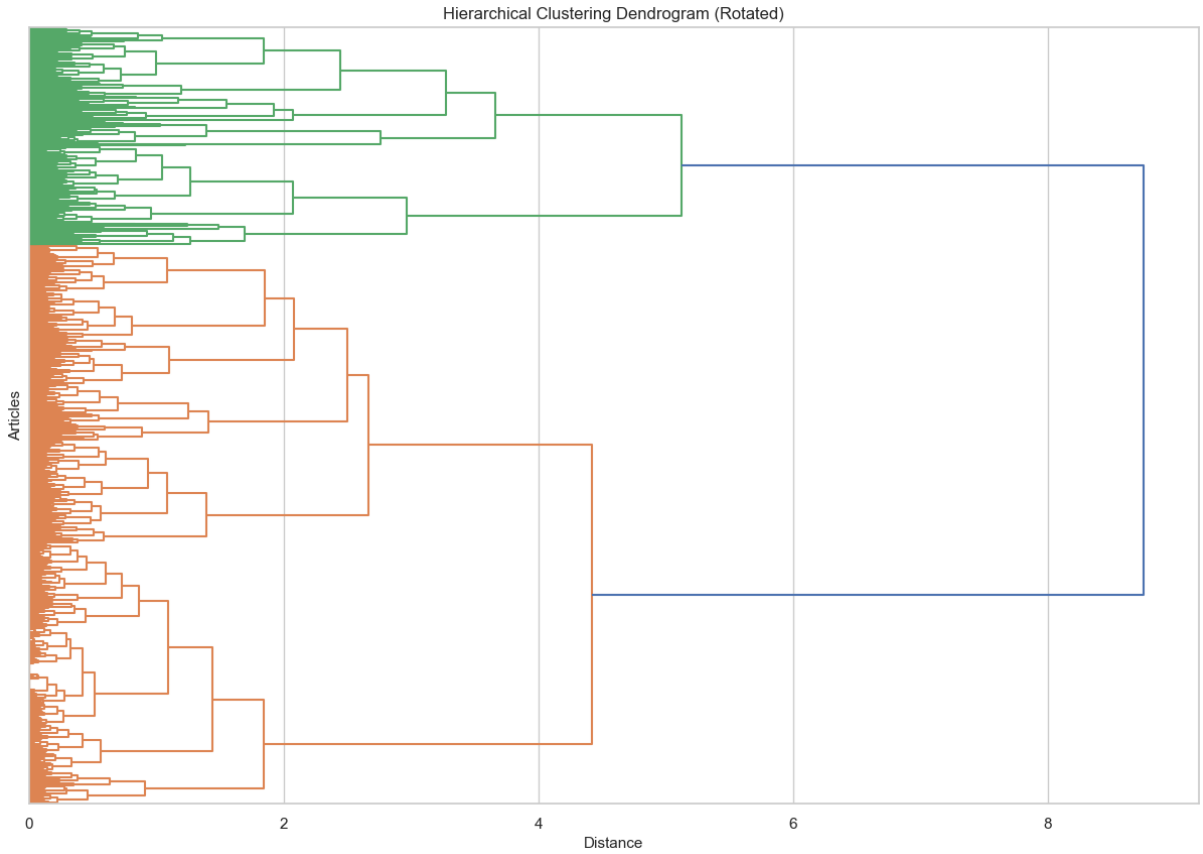

Figure 5.2. Hierarchical Clustering Dendrogram (Rotated)

Figure 5.2 illustrates the Hierarchical Clustering Dendrogram (Rotated), showcasing the hierarchical clustering results for both 2-cluster and 3-cluster solutions.

After implementing the hierarchical clustering algorithm with moral dimensions, the analysis of the 2-cluster solution provided significant insights into the distribution and characteristics of the clusters.

In terms of distribution, Cluster 0 contained 2376 articles, while Cluster 1 had 921 articles.

Cluster 0 focused predominantly on " General Discussions" (51.09%), followed by " General Politics and Social Issues" (30.72%), " National Politics and Parties" (11.41%), and " Israel-Palestine Conflict" (6.73%). Cluster 1's articles were mostly about " General Discussions " (36.48%), " General Politics and Social Issues " (27.14%), " Israel-Palestine Conflict " (19.11%), and " National Politics and Parties " (17.26%).

The moral centroids analysis, represented as units of scaled scores, showed distinct differences between the clusters. Cluster 0 had lower scores across most moral dimensions compared to Cluster 1, with notable differences in "scaled_AuthorityVirtue" (0.106325 for Cluster 0 vs. 0.272364 for Cluster 1) and "scaled_HarmVice" (0.056187 for Cluster 0 vs. 0.195242 for Cluster 1). Cluster 1's higher scores suggest a stronger emphasis on these moral dimensions, indicating a different moral focus compared to Cluster 0.

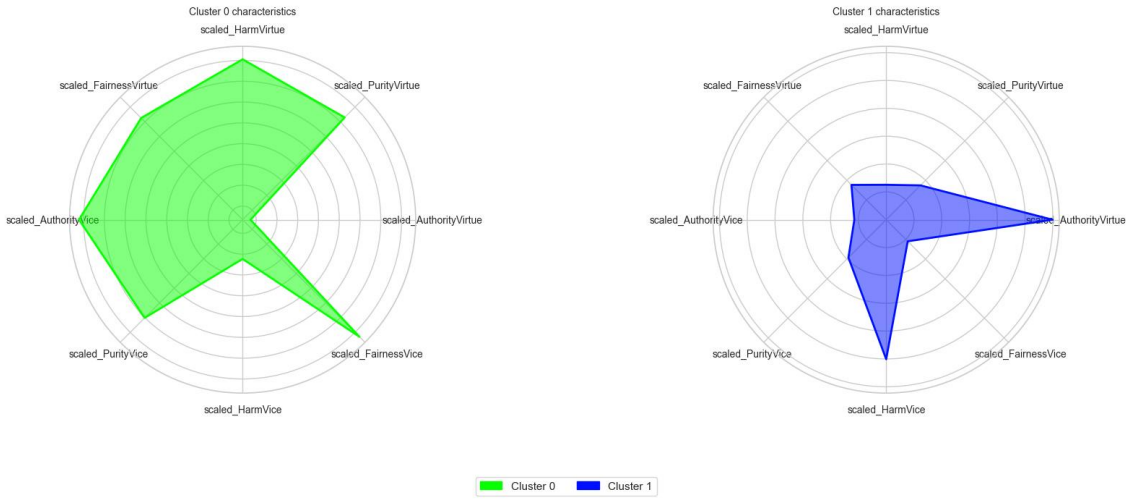

Figure 5.3. Kmeans with Moral Dimensions solution with 2 clusters profile Radar Chart

The radar charts further elucidate these differences. The radar chart for Cluster 0 shows a broader spread across various moral dimensions, indicating a diverse set of moral considerations in the articles associated with this cluster. The most pronounced dimensions for Cluster 0 are Authority Vice and Fairness Virtue, suggesting that the articles in Cluster 0 often emphasize respect for authority and fairness issues. In contrast, the radar chart for Cluster 1 reveals a more concentrated profile. The articles in this cluster show higher values in dimensions such as Authority Virtue and Harm Vice. This concentrated profile indicates a narrower but more intense focus on specific moral dimensions, particularly those related to authority and harm.

The PCA loadings for the 2-cluster solution further elucidate the underlying patterns in the data. For PC1, the top contributing features were "scaled_AuthorityVirtue," "scaled_HarmVice," and "scaled_FairnessVirtue." For PC2, the top contributors were "scaled_HarmVice," "scaled_AuthorityVirtue," and "scaled_PurityVirtue." For PC3, the leading contributors were "scaled_FairnessVirtue," "scaled_FairnessVice," and "scaled_HarmVice."

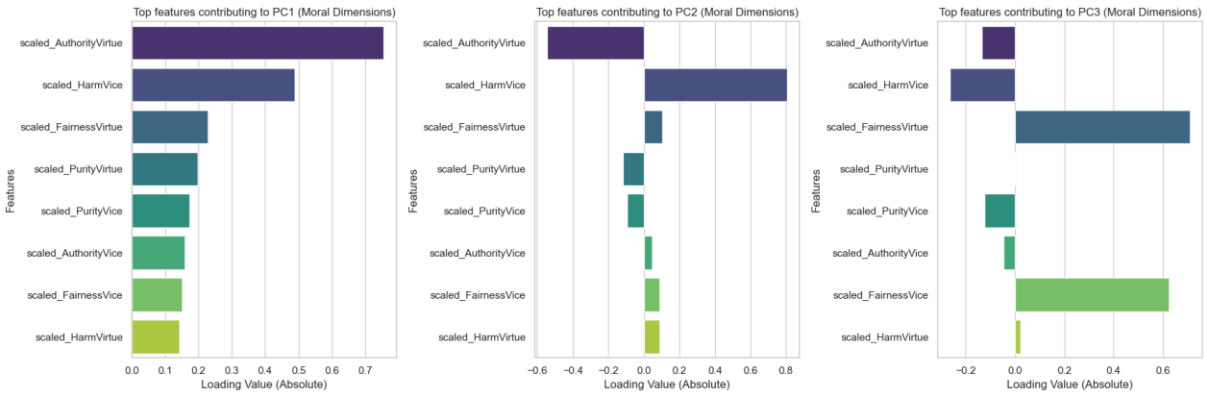

Figure 5.4. Top moral dimensions Moral Dim features contribution per PC

The PCA plot (Figure 5.5) provided a clear visual separation between the two clusters, supporting the numerical findings. This visual separation implies that the articles in each cluster are distinctly different based on their moral dimensions. Furthermore, the percentage of articles in the same cluster by authors depicted the distribution of articles among the top authors, highlighting those with the most significant presence in each cluster.

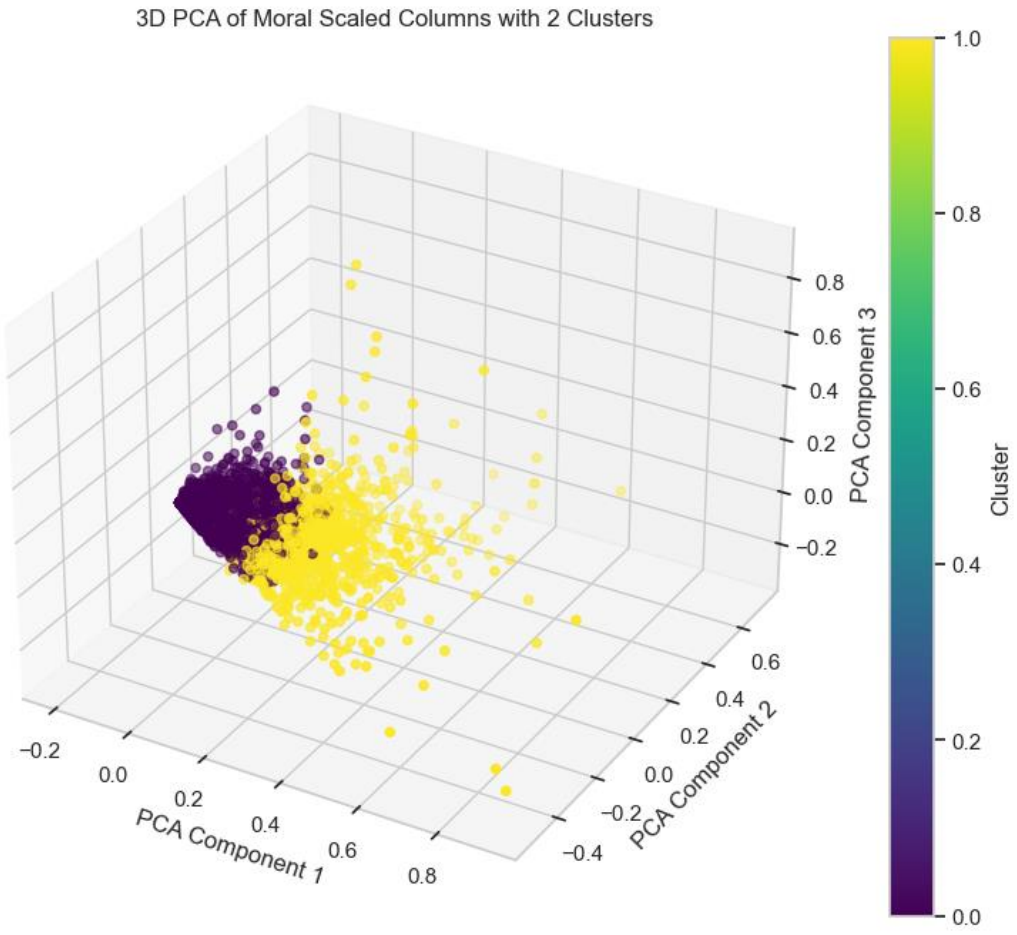

Figure 5.5. 3D PCA plot for Hierarchical Moral scaled Dimensions with 2 Clusters

When examining the number of authors per cluster, Cluster 0 contained 298 authors (83.71%), while Cluster 1 had 58 authors (16.29%). The percentage matrix for the top 10 authors in Cluster 0 included Miguel Esteves Cardoso (198 articles), Henrique Raposo (102 articles), João Miguel Tavares (64 articles), David Pontes (60 articles), and Daniel Oliveira (56 articles). For Cluster 1, the top contributors were António Barreto (25 articles), José Miguel Júdice (24 articles), São José Almeida (24 articles), Teresa de Sousa (24 articles), and Francisco Mendes da Silva (20 articles).

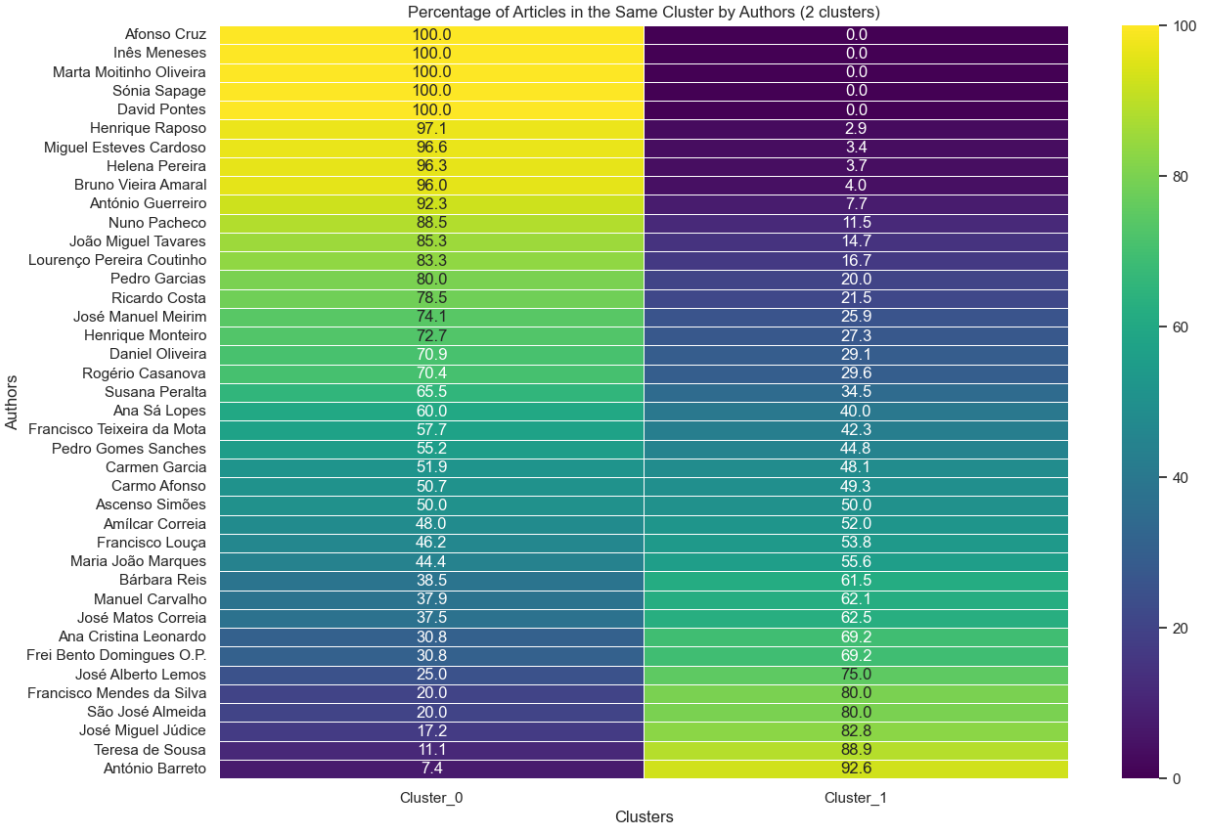

Figure 5.6. Hierarchical on Moral Dimensions Percentage Matrix of Articles in the Same Cluster by Authors on 2 clusters

For summarize we look carefully at our table (Table 5.2) comparison with the most important metrics as following:

Table 5.2. Percentage Topics Cluster 0 and 1 comparison metrics for Hierarchical moral dim with 2 clusters

Percentage of Topics	Cluster 0	Cluster 1
General Discussions	51.09%	36.48%
General Politics and Social Issues	30.72%	27.14%
National Politics and Parties	11.41%	17.26%
Israel-Palestine Conflict	6.73%	19.11%
Climate and Energy	0.04%	

Table 5.3. Moral Centroids Cluster 0 and 1 comparison metrics for Hierarchical moral dim with 2 clusters

Moral Centroids scaled scores	Cluster 0	Cluster 1
scaled_AuthorityVirtue	0.10	0.27
scaled_PurityVirtue	0.03	0.08
scaled_HarmVirtue	0.02	0.05
scaled_FairnessVirtue	0.04	0.09
scaled_AuthorityVice	0.02	0.06
scaled_PurityVice	0.01	0.07
scaled_HarmVice	0.05	0.19
scaled_FairnessVice	0.02	0.05

The findings of our study draw interesting parallels between political discourse in Portugal and the United States. In the US, political conversations often reveal a split between conservatives, who emphasize Authority, Purity, and Loyalty, and liberals, who focus more on Harm and Fairness. Our analysis in Portugal shows a similar division. Cluster 0, reflecting a diverse moral perspective, is akin to US liberals, engaging broadly on social issues and fairness. Conversely, Cluster 1 parallels US conservatives, with a concentrated focus on authority and harm, indicating more intense moral considerations. This moral diversity and intensity have significant implications. Cluster 0 shows moral diversity, suggesting a discourse open to multiple viewpoints. In contrast, Cluster 1 indicates a more focused and intense moral stance, driving stronger, more polarized opinions. Understanding these clusters can guide political strategies. For instance, parties can tailor their messages to resonate with each cluster's moral focus. For Cluster 0, emphasizing social justice, equality, and broad authority issues would be effective. Meanwhile, Cluster 1 would respond better to themes of law and order, national security, and moral integrity. Looking at the political parties and topics, Cluster 0 appeals to broad societal issues and fairness, making it suitable for parties that focus on social justice and equality. On the other hand, Cluster 1 is more appropriate for parties emphasizing authority and harm, with a focus on law and order and security issues. Our study revealed that Cluster 0 was dominant in "General Discussions" (51.09%) and "General Politics and Social Issues" (30.72%), with lower scores across most moral dimensions. Cluster 1, however, concentrated on "General Discussions" (36.48%) and "Israel-Palestine Conflict" (19.11%), showing higher scores in "Authority Virtue" and "Harm Vice.".

6. CONCLUSIONS

The Portuguese political landscape, characterized by its complexity, has been thoroughly examined through various clustering methods, yielding significant insights into the moral dimensions that shape political discourse. Among the methods employed—K-means with Cosine Similarity, K-means with Moral Dimensions, and Hierarchical Clustering with Moral Dimensions—the Hierarchical Clustering with Moral Dimensions emerged as the most robust and insightful framework.

This approach achieved the highest Silhouette Score of 0.3454 and a Davies-Bouldin Index of 1.6471, indicating well-defined and compact clusters. Hierarchical Clustering with Moral Dimensions not only demonstrated a clear separation in moral dimensions but also effectively distinguished the authors within each cluster. This method provided a comprehensive understanding of the moral landscape in Portuguese political discourse, revealing key patterns and distinctions in the moral considerations of different authors and topics.

The findings from the Hierarchical Clustering with Moral Dimensions approach highlight the nuanced moral divisions within Portuguese political discourse. This binary structure mirrors the political landscape in the United States, suggesting a similar division between conservative and liberal moral perspectives in Portugal.

In conclusion, the Hierarchical Clustering with Moral Dimensions approach offers the most robust and insightful framework for analyzing the moral dimensions of Portuguese political discourse. This method provides valuable insights into the complex moral landscape, enabling a deeper understanding of the underlying moral perspectives that shape political narratives in Portugal.

6.1. FUTURE WORK

To build on these findings and enhance our understanding of moral dimensions in Portuguese political discourse, several future research directions are recommended:

Temporal Analysis: Investigate how the moral dimensions and topic distributions evolve over time to understand shifts in political discourse and moral priorities.

Qualitative Deepening: Conduct qualitative case studies of key articles or authors within each cluster to uncover the nuances of their moral reasoning and how it influences political discourse.

Cross-National Comparisons: Compare the findings with those from other countries to identify common patterns or unique aspects of Portuguese political discourse, thereby gaining a broader perspective on moral dimensions in different political contexts.

Advanced Clustering Techniques: Explore advanced clustering algorithms such as DBSCAN or Gaussian Mixture Models to potentially uncover additional layers of complexity within the data.

Sentiment Analysis Integration: Incorporate sentiment analysis to complement the moral dimension analysis, providing deeper insights into the emotional undertones and their impact on political discourse.

Policy and Communication Implications: Analyze how the identified moral clusters influence political communication strategies and policy-making, offering practical insights for political actors and communicators.

This study has laid a solid foundation for understanding the moral dimensions of Portuguese political discourse. Future research will undoubtedly build on these findings, offering even more nuanced insights into the moral underpinnings of political communication and decision-making in Portugal.

BIBLIOGRAPHICAL REFERENCES:

Abdulhai, M., Serapio-Garcia, G., Crepy, C., Valter, D., Canny, J., & Jaques, N. (2023). Moral Foundations of Large Language Models. https://doi.org/10.48550/arXiv.2310.15337

Al-Anzi, F. S., & AbuZeina, D. (2017). Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. Journal of King Saud University - Computer and Information Sciences, 29(2), 189–195. https://doi.org/10.1016/j.jksuci.2016.04.001

Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S. T., & Dehghani, M. (2023). Morality beyond the WEIRD: How the nomological network of morality varies across cultures. Journal of Personality and Social Psychology, 125(5), 1157–1188. https://doi.org/10.1037/pspp0000470

Baskov, O. V., & Noghin, V. D. (2021). The Edgeworth–Pareto Principle in the Case of a Type-2 Fuzzy Preference Relation. Scientific and Technical Information Processing, 48(5), 299–307. https://doi.org/10.3103/S0147688221050014

Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L., & J. Tallón-Ballesteros, A. (2023). The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis. Em P. García Bringas, H. Pérez García, F. J. Martínez de Pisón, F. Martínez Álvarez, A. Troncoso Lora, Á. Herrero, J. L. Calvo Rolle, H. Quintián, & E. Corchado (Eds.), 18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023) (pp. 344–353). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42536-3_33

Chowdhury, R. M. M. I. (2019). The Moral Foundations of Consumer Ethics. Journal of Business Ethics, 158(3), 585–601. https://doi.org/10.1007/s10551-017-3676-2

Cochrane, C., Rheault, L., Godbout, J.-F., Whyte, T., Wong, M. W.-C., & Borwein, S. (2022). The Automatic Analysis of Emotion in Political Speech Based on Transcripts. Political Communication, 39(1), 98–121. https://doi.org/10.1080/10584609.2021.1952497

Demirel, S., Kahraman, E., & Gündüz, U. (2022). A text mining analysis of the change in status of the Hagia Sophia on Twitter: the political discourse and its reflections on the public opinion. Atlantic Journal of Communication, 32(1), 63–90. https://doi.org/10.1080/15456870.2022.2093354

Dickinson, J. L., McLeod, P., Bloomfield, R., & Allred, S. (2016). Which Moral Foundations Predict Willingness to Make Lifestyle Changes to Avert Climate Change in the USA? PLOS ONE, 11(10), e0163852. https://doi.org/10.1371/journal.pone.0163852

Feature scaling. (2024). Em Wikipedia. https://en.wikipedia.org/w/index.php?title=Feature_scaling&oldid=1228410225

Gao, X., Ding, X., Han, T., & Kang, Y. (2022). Analysis of influencing factors on excellent teachers' professional growth based on DB-Kmeans method. EURASIP Journal on Advances in Signal Processing, 2022(1), 117. https://doi.org/10.1186/s13634-022-00948-2

Gao, X., Ding, X., Wang, W., Wang, G., Kang, Y., & Wang, S. (2022). Keywords Clustering for the Interview Texts Based on Kmeans Algorithm. Em Q. Liang, W. Wang, J. Mu, X. Liu, & Z. Na (Eds.), Artificial Intelligence in China (pp. 600–606). Springer. https://doi.org/10.1007/978-981-16-9423-3_75

Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., & Dehghani, M. (2016). Morality Between the Lines: Detecting Moral Sentiment In Text. https://media.mola-lab.org/file/1702633610480-MoralityBetweentheLinesDetectingMoralSentimentinText.pdf

González-Santos, C., Vega-Rodríguez, M. A., Pérez, C. J., López-Muñoz, J. M., & Martínez-Sarriegui, I. (2023). Automatic assignment of moral foundations to movies by word embedding. Knowledge-Based Systems, 270, 110539. https://doi.org/10.1016/j.knosys.2023.110539

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2012). Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism (SSRN Scholarly Paper 2184440). https://papers.ssrn.com/abstract=2184440

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. Em P. Devine & A. Plant (Eds.), Advances in Experimental Social Psychology (Vol. 47, pp. 55–130). Academic Press. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology, 96(5), 1029–1046. https://doi.org/10.1037/a0015141

Haidt, J. (2008). Morality. Perspectives on Psychological Science, 3(1), 65–72. https://doi.org/10.1111/j.1745-6916.2008.00063.x

Haidt, J. (2012). The Righteous Mind: Why Good People Are Divided by Politics and Religion. Knopf Doubleday Publishing Group.

Haidt, J., & Graham, J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. Social Justice Research, 20(1), 98–116. https://doi.org/10.1007/s11211-007-0034-z

Haidt, J., Graham, J., & Joseph, C. (2009). Above and Below Left–Right: Ideological Narratives and Moral Foundations. Psychological Inquiry, 20(2–3), 110–119. https://doi.org/10.1080/10478400903028573

Haidt, J., & Joseph, C. (2004). Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. Daedalus, 133(4), 55–66. https://doi.org/10.1162/0011526042365555

Hauke, J., & Kossowski, T. (2011). Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. Quaestiones Geographicae, 30(2), 87–93. https://doi.org/10.2478/v10117-011-0021-1

Henderi, H., Wahyuningsih, T., & Rahwanto, E. (2021). Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. International Journal of Informatics and Information Systems, 4(1), Artigo 1. https://doi.org/10.47738/ijiis.v4i1.73

Hopp, F. R., Amir, O., Fisher, J. T., Grafton, S., Sinnott-Armstrong, W., & Weber, R. (2023). Moral foundations elicit shared and dissociable cortical activation modulated by political ideology. Nature Human Behaviour, 7(12), 2182–2198. https://doi.org/10.1038/s41562-023-01693-8

Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (2021). The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. Behavior Research Methods, 53(1), 232–246. https://doi.org/10.3758/s13428-020-01433-0

Hu, N., Ho, K. C., & Fan, P. S. (2024). Malaysian Chinese folk beliefs on Facebook based on LDA topic modelling. Humanities and Social Sciences Communications, 11(1), 547. https://doi.org/10.1057/s41599-024-03066-6

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Joseph, C. M., Graham, J., & Haidt, J. (2009). The End of Equipotentiality: A Moral Foundations Approach to Ideology-Attitude Links and Cognitive Complexity. Psychological Inquiry, 20(2–3), 172–176. https://doi.org/10.1080/10478400903088882

Kaur, R., & Sasahara, K. (2016). Quantifying moral foundations from various topics on Twitter conversations. 2016 IEEE International Conference on Big Data (Big Data), 2505–2512. https://doi.org/10.1109/BigData.2016.7840889

Kiran, A., & Vasumathi, D. (2020). Data Mining: Min–Max Normalization Based Data Perturbation Technique for Privacy Preservation. Em K. S. Raju, A. Govardhan, B. P. Rani, R. Sridevi, & M. R. Murty (Eds.), Proceedings of the Third International Conference on Computational Intelligence and Informatics (pp. 723–734). Springer. https://doi.org/10.1007/978-981-15-1480-7_66

Lee, J. H., & Ostwald, M. J. (2024). Latent Dirichlet Allocation (LDA) topic models for Space Syntax studies on spatial experience. City, Territory and Architecture, 11(1), 3. https://doi.org/10.1186/s40410-023-00223-3

Lenssen, L., & Schubert, E. (2022). Clustering by Direct Optimization of the Medoid Silhouette. Em T. Skopal, F. Falchi, J. Lokoč, M. L. Sapino, I. Bartolini, & M. Patella (Eds.), Similarity Search and Applications (pp. 190–204). Springer International Publishing. https://doi.org/10.1007/978-3-031-17849-8_15

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space (arXiv:1301.3781). arXiv. https://doi.org/10.48550/arXiv.1301.3781

Mills, B., & Wilner, A. (2023). The science behind “values”: Applying moral foundations theory to strategic foresight. FUTURES & FORESIGHT SCIENCE, 5(1), e145. https://doi.org/10.1002/ffo2.145

Moral Foundations Theory | moralfoundations.org. (2023). Obtido 14 de setembro de 2023, de https://moralfoundations.org/

Mulyani, H., Setiawan, R. A., & Fathi, H. (2023). Optimization of K Value in Clustering Using Silhouette Score (Case Study: Mall Customers Data). Journal of Information Technology and Its Utilization, 6(2), Artigo 2. https://doi.org/10.56873/jitu.6.2.5243

Musschenga, B. (2013). The promises of moral foundations theory. Journal of Moral Education, 42(3), 330–345. https://doi.org/10.1080/03057240.2013.817326

Naghizadeh, A., & Metaxas, D. N. (2020). Condensed Silhouette: An Optimized Filtering Process for Cluster Selection in K-Means. Procedia Computer Science, 176, 205–214. https://doi.org/10.1016/j.procs.2020.08.022

Parks, L., & Peters, W. (2023). Natural Language Processing in Mixed-methods Text Analysis: A Workflow Approach. International Journal of Social Research Methodology, 26(4), 377–389. https://doi.org/10.1080/13645579.2021.2018905

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543. https://doi.org/10.3115/v1/D14-1162

Ros, F., Riad, R., & Guillaume, S. (2023). PDBI: A partitioning Davies-Bouldin index for clustering evaluation. Neurocomputing, 528, 178–199. https://doi.org/10.1016/j.neucom.2023.01.043

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sauer, H. (2015). Can’t We All Disagree More Constructively? Moral Foundations, Moral Reasoning, and Political Disagreement. Neuroethics, 8(2), 153–169. https://doi.org/10.1007/s12152-015-9235-6

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP Journal on Wireless Communications and Networking, 2021(1), 31. https://doi.org/10.1186/s13638-021-01910-w

Shutaywi, M., & Kachouie, N. N. (2021). Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. Entropy, 23(6), Artigo 6. https://doi.org/10.3390/e23060759

Simpson, A. (2017). Moral Foundations Theory. In: Zeigler-Hill, V., Shackelford, T. (eds) Encyclopedia of Personality and Individual Differences. Springer, Cham. https://doi.org/10.1007/978-3-319-28099-8_1253-1

Sinsomboonthong, S. (2022). Performance Comparison of New Adjusted Min-Max with Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network Classification. International Journal of Mathematics and Mathematical Sciences, 2022(1), 3584406. https://doi.org/10.1155/2022/3584406

Sitompul, B. J. D., Sitompul, O. S., & Sihombing, P. (2019). Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm. Journal of Physics: Conference Series, 1235(1), 012015. https://doi.org/10.1088/1742-6596/1235/1/012015

Soares, V. H. A., Campello, R. J. G. B., Nourashrafeddin, S., Milios, E., & Naldi, M. C. (2019). Combining semantic and term frequency similarities for text clustering. Knowledge and Information Systems, 61(3), 1485–1516. https://doi.org/10.1007/s10115-018-1278-7

Stewart, B. D., & Morris, D. S. M. (2021). Moving Morality Beyond the In-Group: Liberals and Conservatives Show Differences on Group-Framed Moral Foundations and These Differences Mediate the Relationships to Perceived Bias and Threat. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.579908

Suhler, C. L., & Churchland, P. (2011). Can Innate, Modular "Foundations" Explain Morality? Challenges for Haidt's Moral Foundations Theory. Journal of Cognitive Neuroscience, 23(9), 2103–2116. https://doi.org/10.1162/jocn.2011.21637

Takikawa, H., & Sakamoto, T. (2020). The moral–emotional foundations of political discourse: A comparative analysis of the speech records of the U.S. and the Japanese legislatures.

Quality & Quantity: International Journal of Methodology, 54(2), 547–566. https://doi.org/10.1007/s11135-019-00912-7

Thakkar, A., & Chaudhari, K. (2020). Predicting stock trend using an integrated term frequency–inverse document frequency-based feature weight matrix with neural networks. Applied Soft Computing, 96, 106684. https://doi.org/10.1016/j.asoc.2020.106684

The Righteous Mind. (sem data). The Righteous Mind. Obtido 19 de setembro de 2023, de https://righteousmind.com/

Theocharis, Y., & Jungherr, A. (2021). Computational Social Science and the Study of Political Communication. Political Communication, 38(1–2), 1–22. https://doi.org/10.1080/10584609.2020.1833121

Vafaei, N., Ribeiro, R. A., & Camarinha-Matos, L. M. (2020). Selecting Normalization Techniques for the Analytical Hierarchy Process. Em L. M. Camarinha-Matos, N. Farhadi, F. Lopes, & H. Pereira (Eds.), Technological Innovation for Life Improvement (pp. 43–52). Springer International Publishing. https://doi.org/10.1007/978-3-030-45124-0_4

Vichi, M., Cavicchia, C., & Groenen, P. J. F. (2022). Hierarchical Means Clustering. Journal of Classification, 39(3), 553–577. https://doi.org/10.1007/s00357-022-09419-7

Wilkerson, J., & Casas, A. (2017). Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. Annual Review of Political Science, 20(1), 529–544. https://doi.org/10.1146/annurev-polisci-052615-025542

Yu, D., & Xiang, B. (2023). Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling. Expert Systems with Applications, 225, 120114. https://doi.org/10.1016/j.eswa.2023.120114

Zúquete, M., Orghian, D., & Pinheiro, F. L. (2023, June). A Moral Foundations Dictionary for the European Portuguese Language: The Case of Portuguese Parliamentary Debates. In International Conference on Computational Science (pp. 421-434). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35995-8_30

ANNEXES

Variables descriptions

Date: The publication date of the content.

Time: The time of publication.

Political Party: The political affiliation mentioned, if any.

Topic Title: The title of the news or opinion piece.

News Content: The main content of the article.

Link: The URL where the content was published.

Author: The name of the content's author.

Type of Content: Categories such as "Opinião", "Exclusivo Análise", and "Megafone".

Editor: The publication, either "Publico" or "Expresso".

Number of Shares: The count of how many times the content was shared (only for Publico).

Number of Comments: The count of comments (only for Publico).

Week Name: The name of the week in which the content was published.

Week Number: The numerical representation of the week.

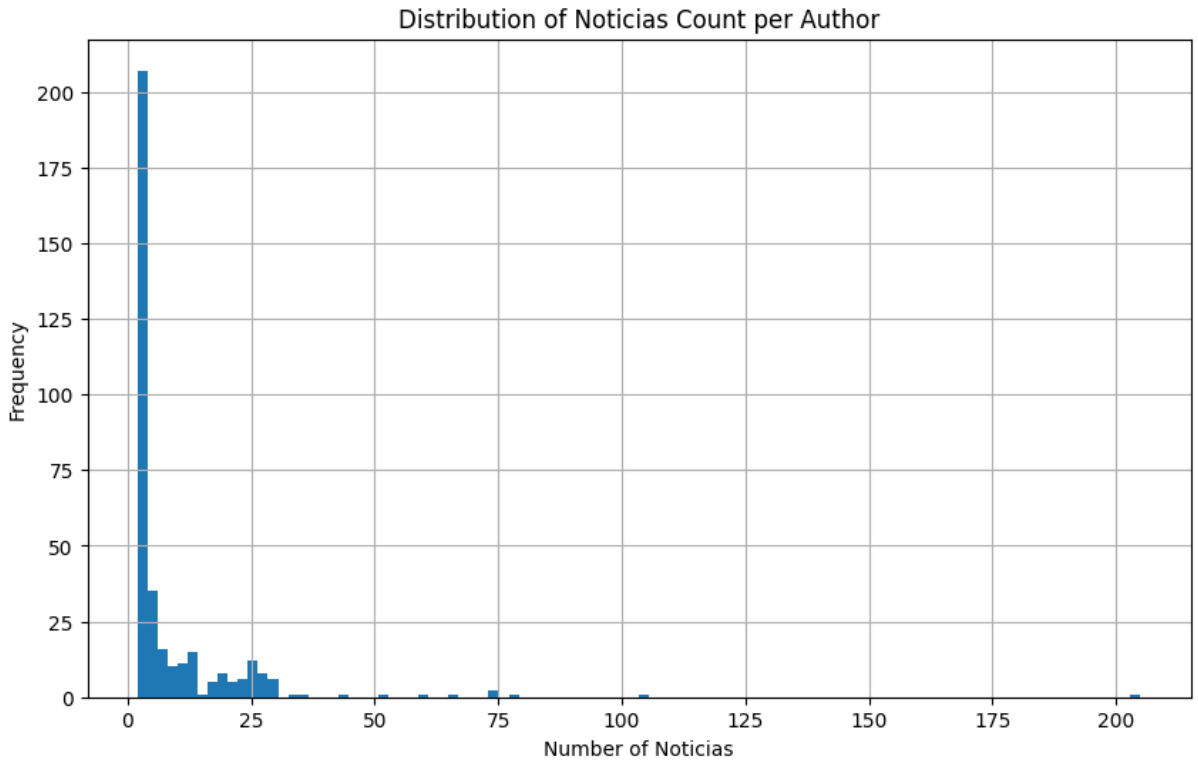

Figure A.1. Histogram for column Noticia per author

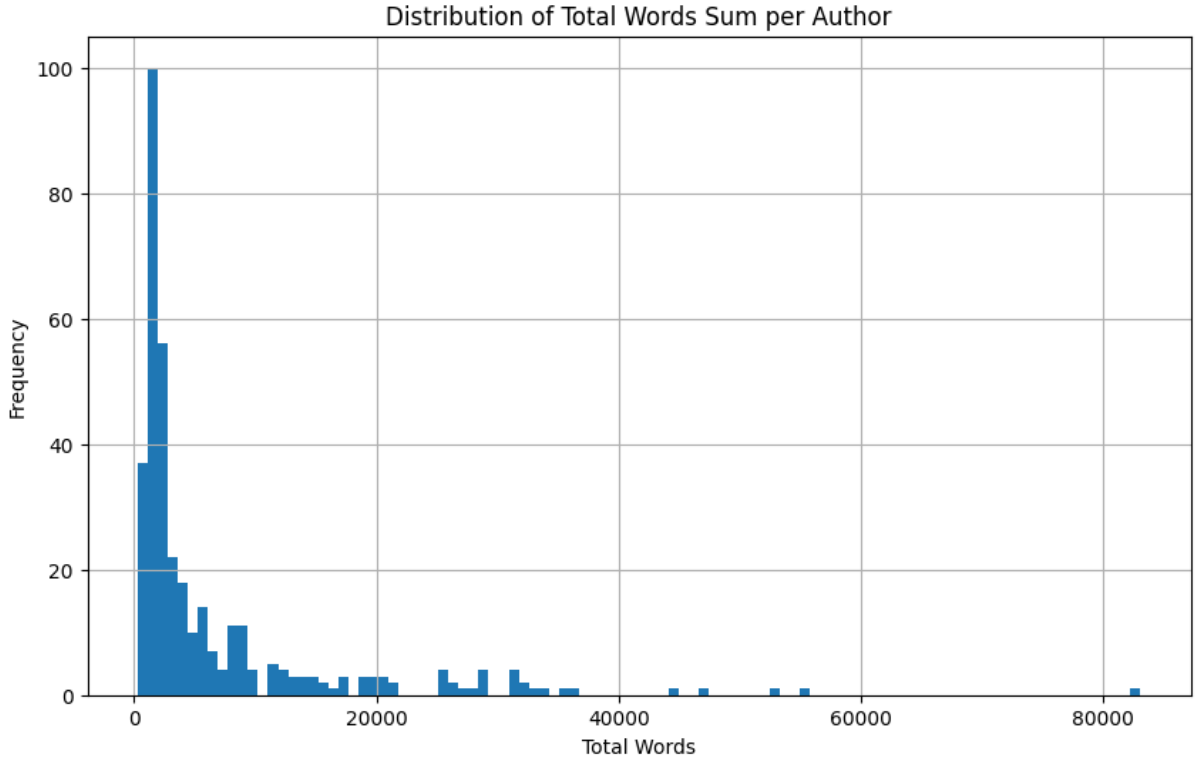

Figure A.2. Histogram for column Total Words sum per author

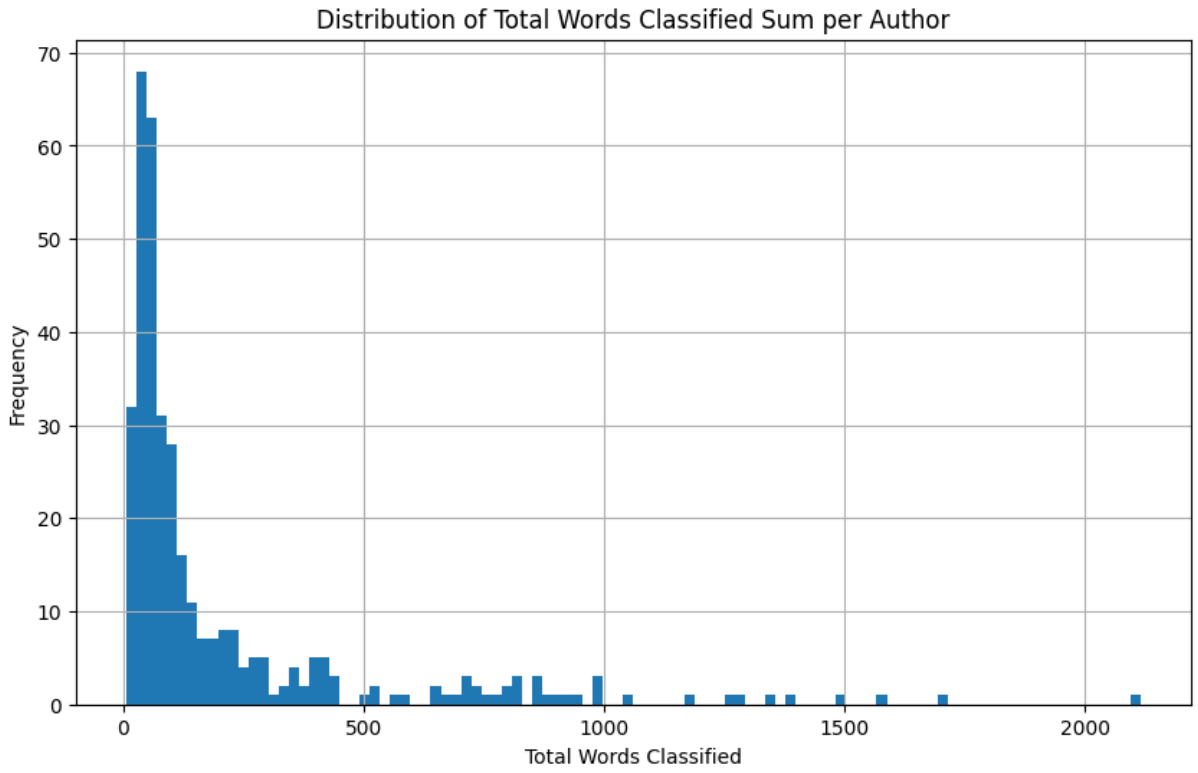

Figure A. 3. Histogram for column Total Words Classified

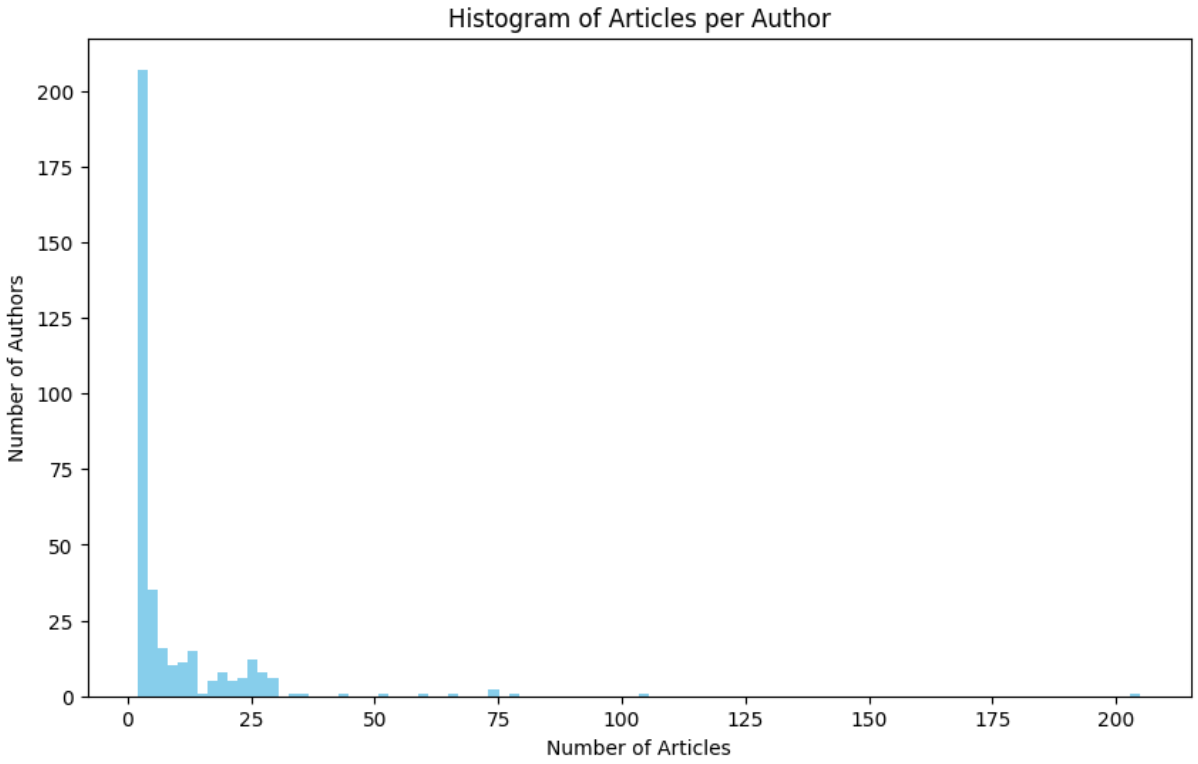

Figure A.4. Histogram for column Articles

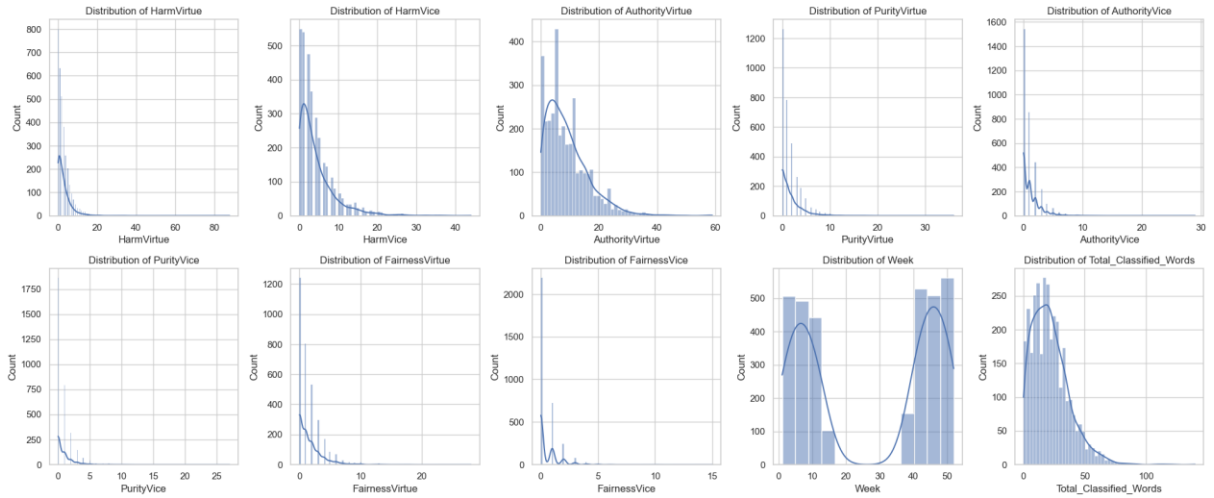

Figure A.5. Histogram for all numeric columns

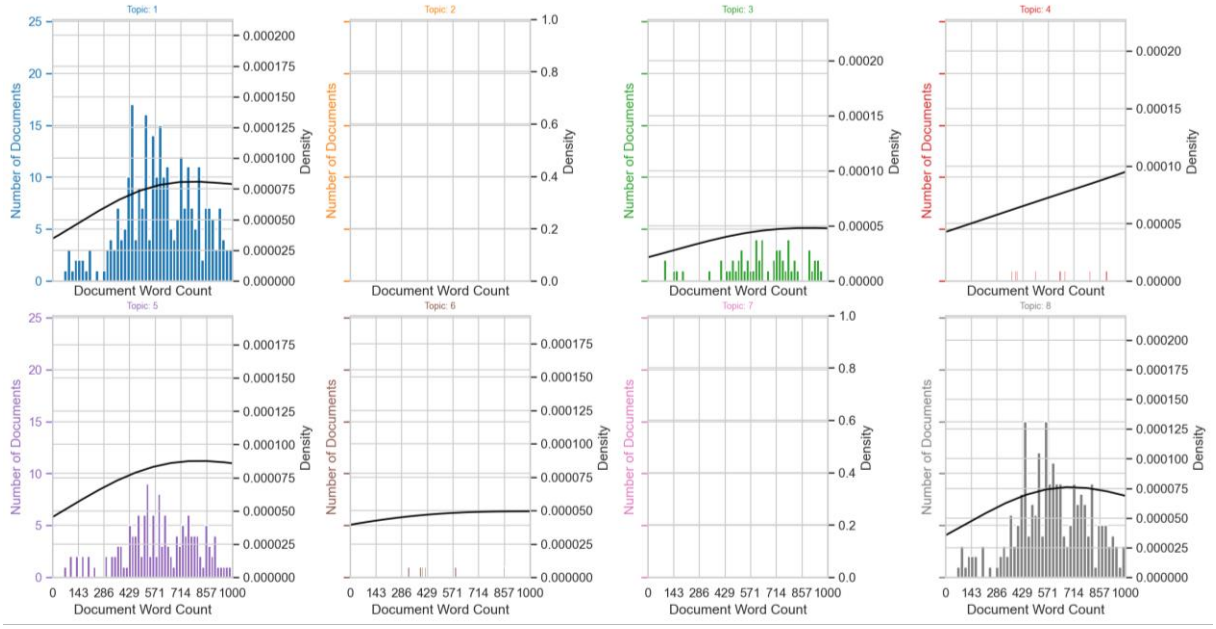

Figure A.6. Distribution of document word counts by dominant topic

Figure A.7. Word clouds for the eight topics identified by the LDA model

		Classified Noticia	Count of Noticias	Sum of Classified Words	Sum of Total Words	Percentage of Total Words	Percentage of Classified Words	Percentage of Noticias	Percentage of Classified Noticias
Author	**Topic_name**								
Miguel Esteves Cardoso	General Discussions	193	199	1567	79924	3.360830	2.165621	6.035790	5.940289
Henrique Raposo	General Discussions	66	70	620	21621	0.909170	0.856851	2.123142	2.031394
Afonso Cruz	General Discussions	51	51	215	11495	0.483368	0.297134	1.546861	1.569714
Ricardo Costa	General Politics and Social Issues	35	36	817	23517	0.988897	1.129108	1.091902	1.077255
João Miguel Tavares	General Discussions	29	29	508	17141	0.720785	0.702065	0.879588	0.892582
...	...	...	...	...	...	...	...	...	...
Miguel Raimundo	General Discussions	1	1	6	405	0.017030	0.008292	0.030331	0.030779
	General Politics and Social Issues	1	1	26	549	0.023086	0.035932	0.030331	0.030779
José Manuel Meirim	Israel-Palestine Conflict	1	1	16	444	0.018670	0.022112	0.030331	0.030779
António Félix	General Politics and Social Issues	1	1	25	638	0.026828	0.034550	0.030331	0.030779
Eduardo Marçal Grilo	Israel-Palestine Conflict	1	1	52	1356	0.057020	0.071865	0.030331	0.030779

Figure A.8. Figure 0.8. Pivot table Author, Topic Name

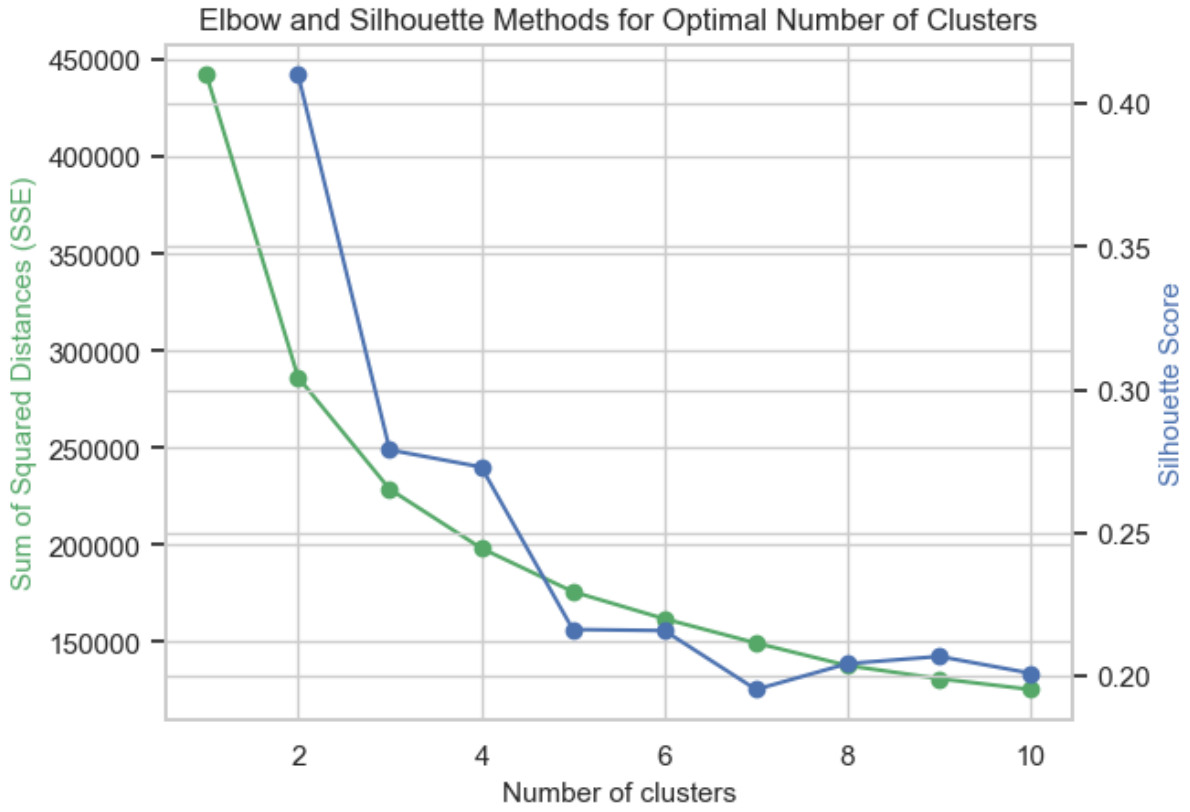

Figure A.9. Elbow and Silhouette Methods Kmeans Cluster Solution with CS for Optimal Number of Clusters

To begin our analysis, we employed the K-means algorithm with cosine similarity. Using the Elbow Method and the Silhouette Method, we determined that the optimal number of clusters was two. The resulting two-cluster solution revealed distinct thematic and authorial distributions.

Cluster 0, containing 2,568 articles, demonstrated a broad focus on various topics. The primary themes included "General Discussions" (43.50%), "General Politics and Social Issues" (32.20%), "National Politics and Parties" (14.84%), "Israel-Palestine Conflict" (9.42%), and "Climate and Energy" (0.04%). This cluster was represented by 333 authors, accounting for 93.54% of the total. Notable contributors included Miguel Esteves Cardoso (104 articles), Carmo Afonso (67 articles), Henrique Raposo (65 articles), João Miguel Tavares (56 articles), and Ricardo Costa (51 articles).

Conversely, Cluster 1 contained 729 articles with a more focused thematic distribution. The dominant themes were "General Discussions" (59.40%), followed by "General Politics and Social Issues" (20.99%), "Israel-Palestine Conflict" (12.89%), and "National Politics and Parties" (6.72%). This cluster comprised 23 authors, representing 6.46% of the total. Key contributors were Gonçalo M. Tavares (12 articles), André Barata (6 articles), Carla Quevedo (5 articles), João Duque (5 articles), and Tiago Luz Pedro (5 articles).

The moral centroid analysis revealed significant differences between the clusters. Cluster 0 had higher scores across almost all moral dimensions compared to Cluster 1, with notable differences in "Authority Virtue" (0.182276 for Cluster 0 vs. 0.048546 for Cluster 1) and "Fairness Vice" (0.094980 for Cluster 0 vs. 0.095211 for Cluster 1).

The clustering effectively separated the authors. For example, Miguel Esteves Cardoso had 50.7% of his articles in Cluster 0 and 49.3% in Cluster 1. In contrast, authors like São José Almeida and Teresa de Sousa had 100% of their articles in Cluster 0, showing a clear distinction in their clustering.

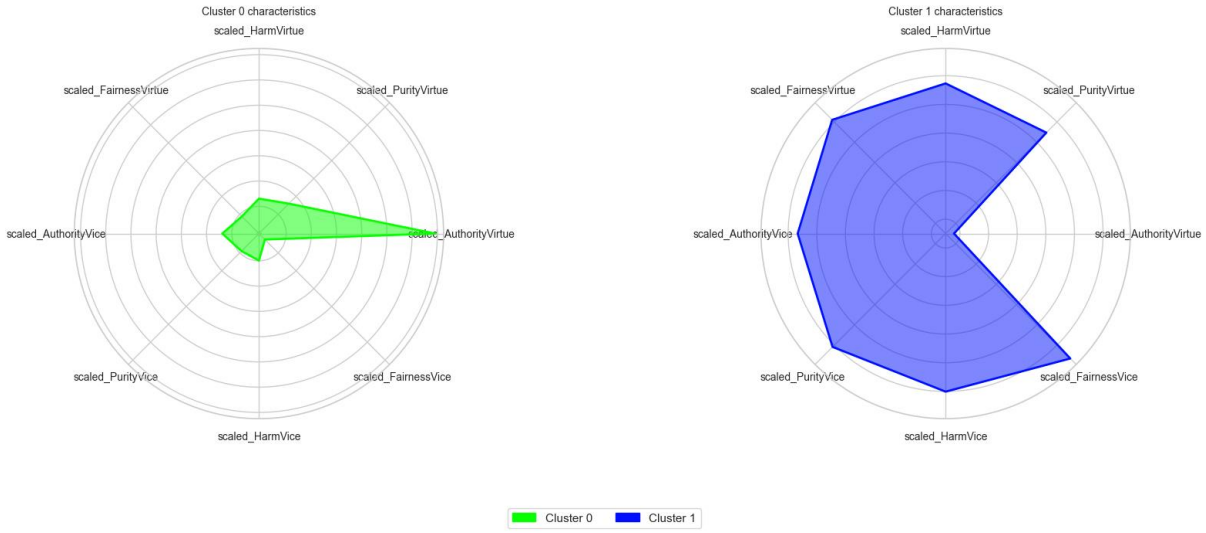

Figure A.10. Kmeans Cluster Solution, Cosine Similarity profile Radar Chart

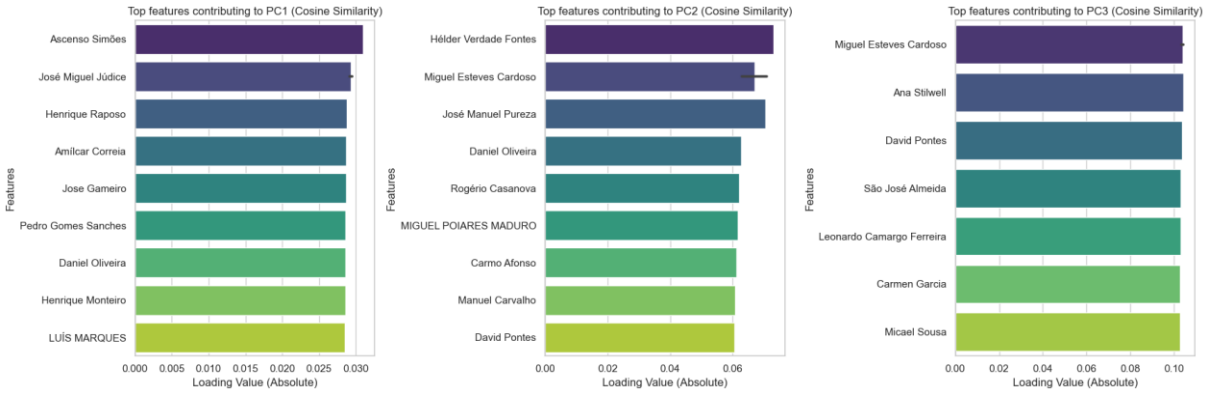

Figure A.11. Two Clusters, Kmeans Cluster Solution, Cosine Similarity 3D top features Author's contribution per PC

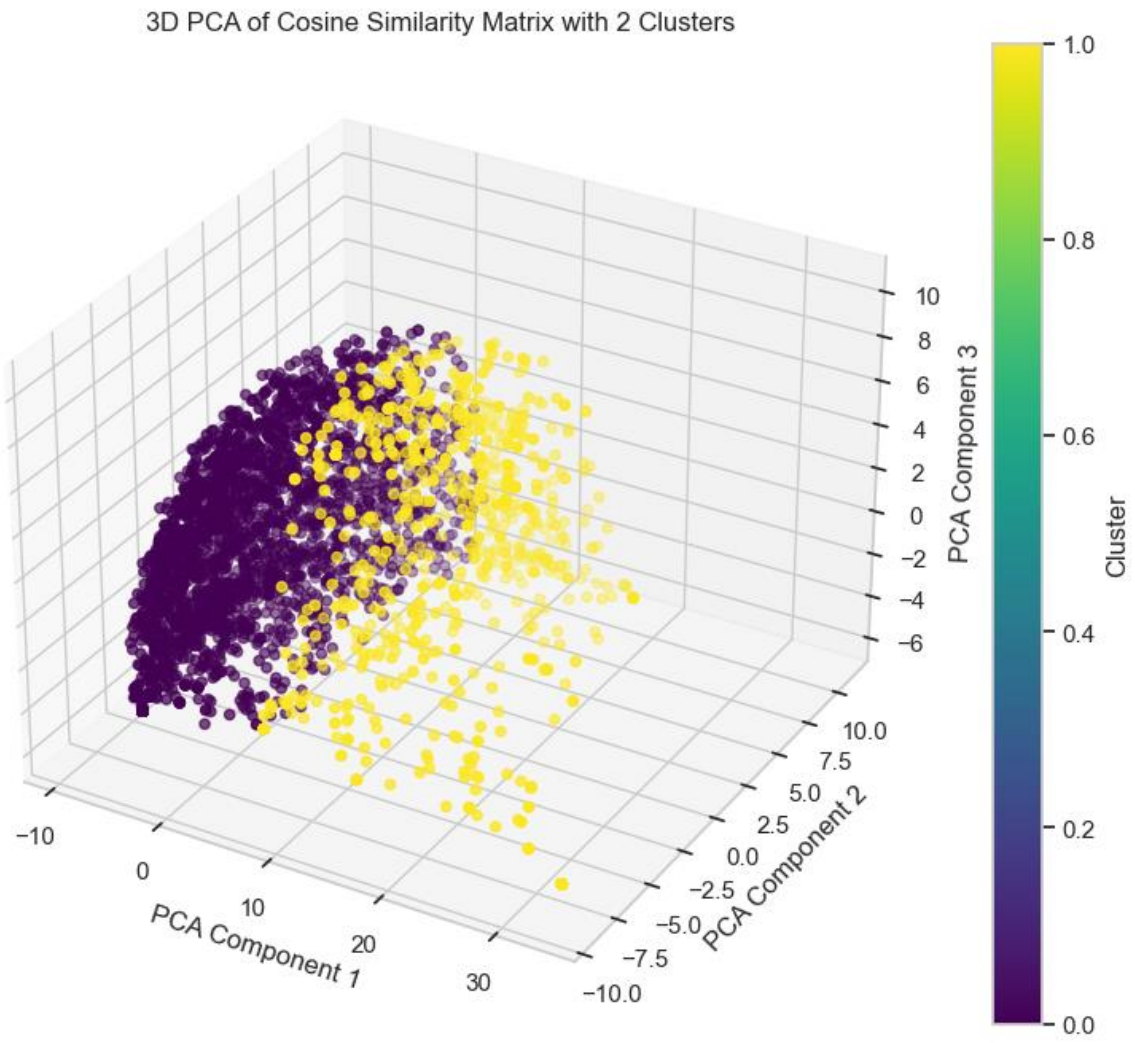

Figure A.12. Two Clusters, Kmeans Cluster Solution, Cosine Similarity 3D PCA

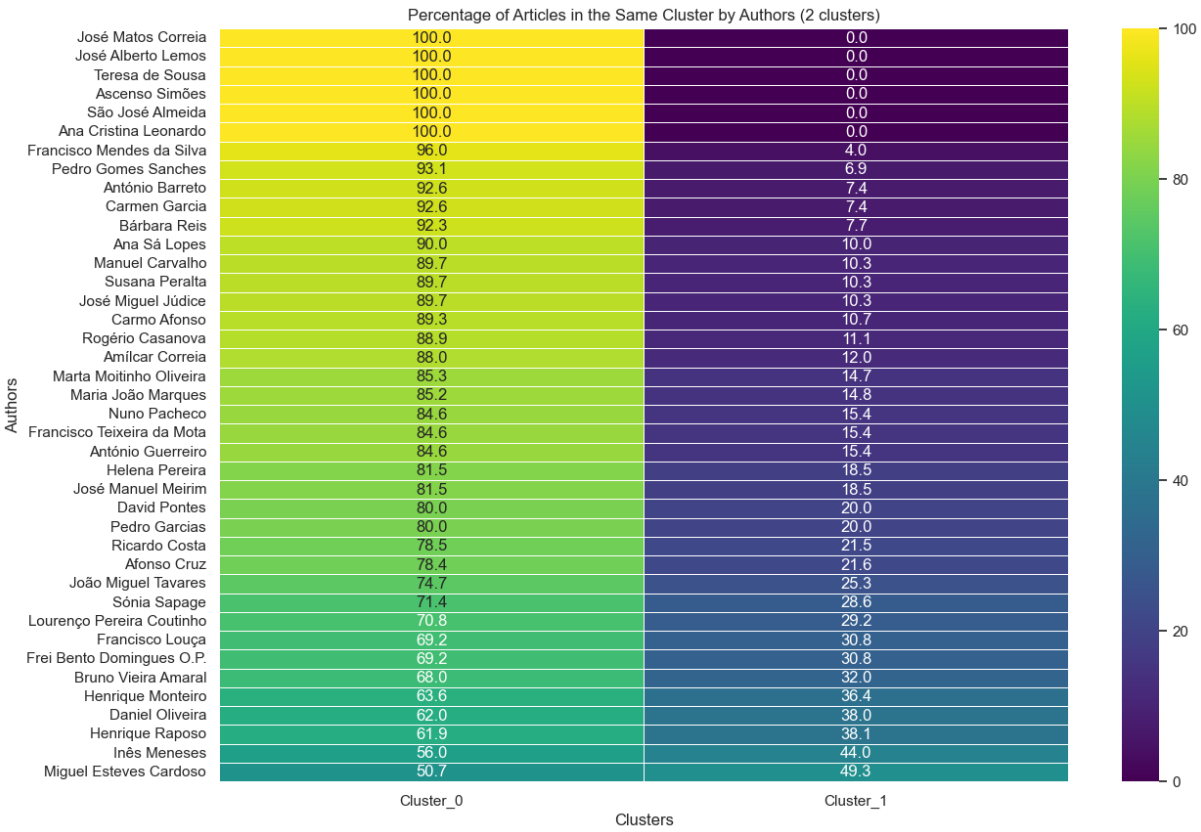

Figure A. 13. Two Clusters, Kmeans Cluster Solution, Cosine Similarity Percentage Matrix of Articles in the Same Cluster by Authors

Table A.1. Cluster 0 and 1 comparison metrics Two Clusters, Kmeans Cluster Solution Cosine Similarity

Metric	Cluster 0	Cluster 1
Silhouette Score	0.41	
Davies-Bouldin Index	1.2214	
Number of Articles	2568	729
Number of Authors	333	23
Percentage of Authors	93.54%	6.46%
Magnitude of Cluster Centroid	0.22	0.14
Percentage of Topics		
General Discussions	43.50%	59.40%
General Politics and Social Issues	32.20%	20.99%
National Politics and Parties	14.84%	6.72%
Israel-Palestine Conflict	9.42%	12.89%
Climate and Energy	0.04%	0.00%
Moral Centroids scaled scores		
scaled_AuthorityVirtue	0.18	0.04

Metric	Cluster 0	Cluster 1
scaled_PurityVirtue	0.04	0.03
scaled_HarmVirtue	0.04	0.02
scaled_FairnessVirtue	0.05	0.05
scaled_AuthorityVice	0.04	0.03
scaled_PurityVice	0.03	0.03
scaled_HarmVice	0.09	0.09
scaled_FairnessVice	0.03	0.04

Figure A.14. Kmeans Cluster Solution, Cosine Similarity Clusters Characteristics Bar chart

Cluster 0 contained 2,376 articles, primarily focused on "General Discussions" (51.09%), "General Politics and Social Issues" (30.72%), "National Politics and Parties" (11.41%), and "Israel-Palestine Conflict" (6.73%). This cluster included 298 authors, accounting for 83.71% of the total, with top contributors being Miguel Esteves Cardoso (198 articles), Henrique

Raposo (102 articles), João Miguel Tavares (64 articles), David Pontes (60 articles), and Daniel Oliveira (56 articles).

Cluster 1 contained 921 articles, focusing on "General Discussions" (36.48%), "General Politics and Social Issues" (27.14%), "Israel-Palestine Conflict" (19.11%), and "National Politics and Parties" (17.26%). This cluster included 58 authors, representing 16.29% of the total, with the top contributors being António Barreto (25 articles), José Miguel Júdice (24 articles), São José Almeida (24 articles), Teresa de Sousa (24 articles), and Francisco Mendes da Silva (20 articles).

The moral centroids analysis revealed significant differences between the clusters. Cluster 0 had lower scores across most moral dimensions compared to Cluster 1, with notable differences in "Authority Virtue" (0.106325 for Cluster 0 vs. 0.272364 for Cluster 1) and "Harm Vice" (0.056187 for Cluster 0 vs. 0.195242 for Cluster 1).

The clustering also effectively separated the authors. For example, Miguel Esteves Cardoso had 96.6% of his articles in Cluster 0 and 3.4% in Cluster 1. In contrast, authors like São José Almeida and Teresa de Sousa had 80.0% and 88.9% of their articles in Cluster 1, respectively, showing a clear distinction in their clustering.

The results from all three clustering approaches highlight significant distinctions in the distribution of articles, topics, and moral dimensions between the clusters. The K-means with Cosine Similarity solution revealed a broad range of topics and a high number of authors in Cluster 0, while Cluster 1 was more focused and had fewer authors. The moral centroid analysis indicated a more diverse set of moral considerations in Cluster 0 compared to Cluster 1.

The K-means with Moral Dimensions solution provided a different perspective, showing a clear separation in moral focus between the clusters. Cluster 0 included a majority of the articles and authors, with a wide range of moral dimensions, while Cluster 1 had a more concentrated moral focus with higher scores in specific moral dimensions such as "Authority Virtue" and "Harm Vice."

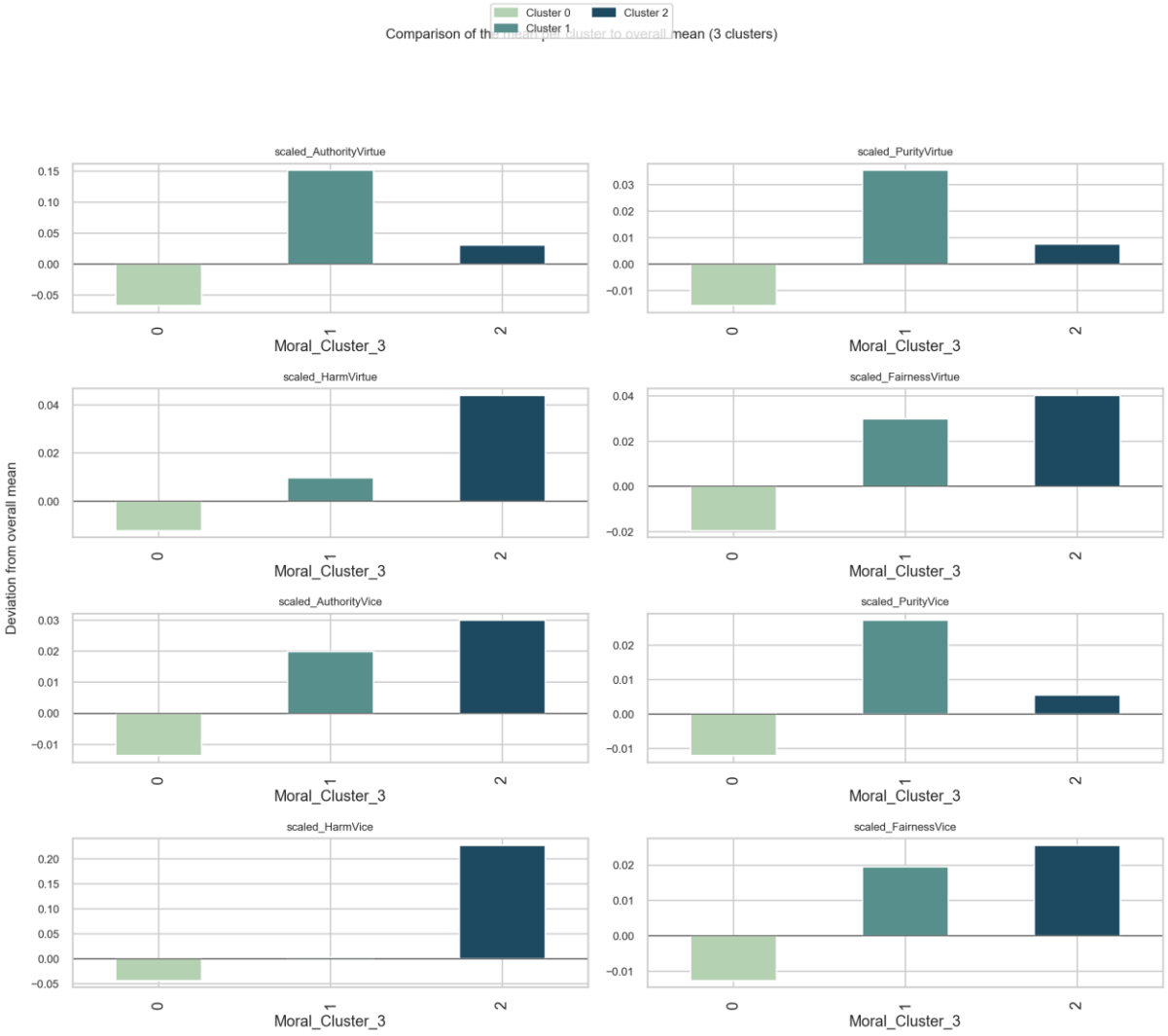

Figure A.15. Three Clusters, Kmeans Cluster Solution, Cosine Similarity Clusters Characteristics Bar Chart

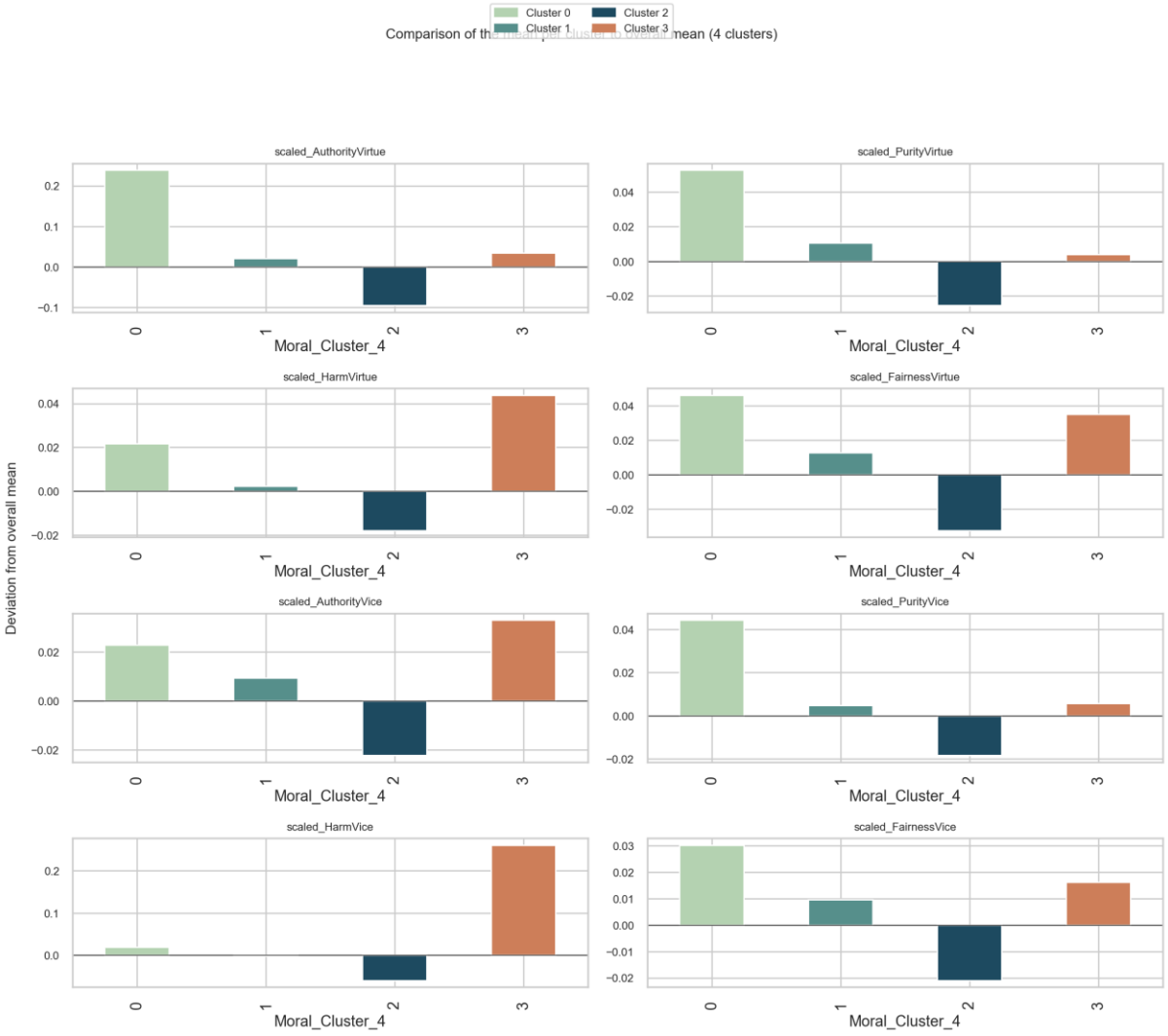

Figure A.16. Four Clusters, Kmeans Cluster Solution Cosine Similarity, Clusters Characteristics Bar Chart

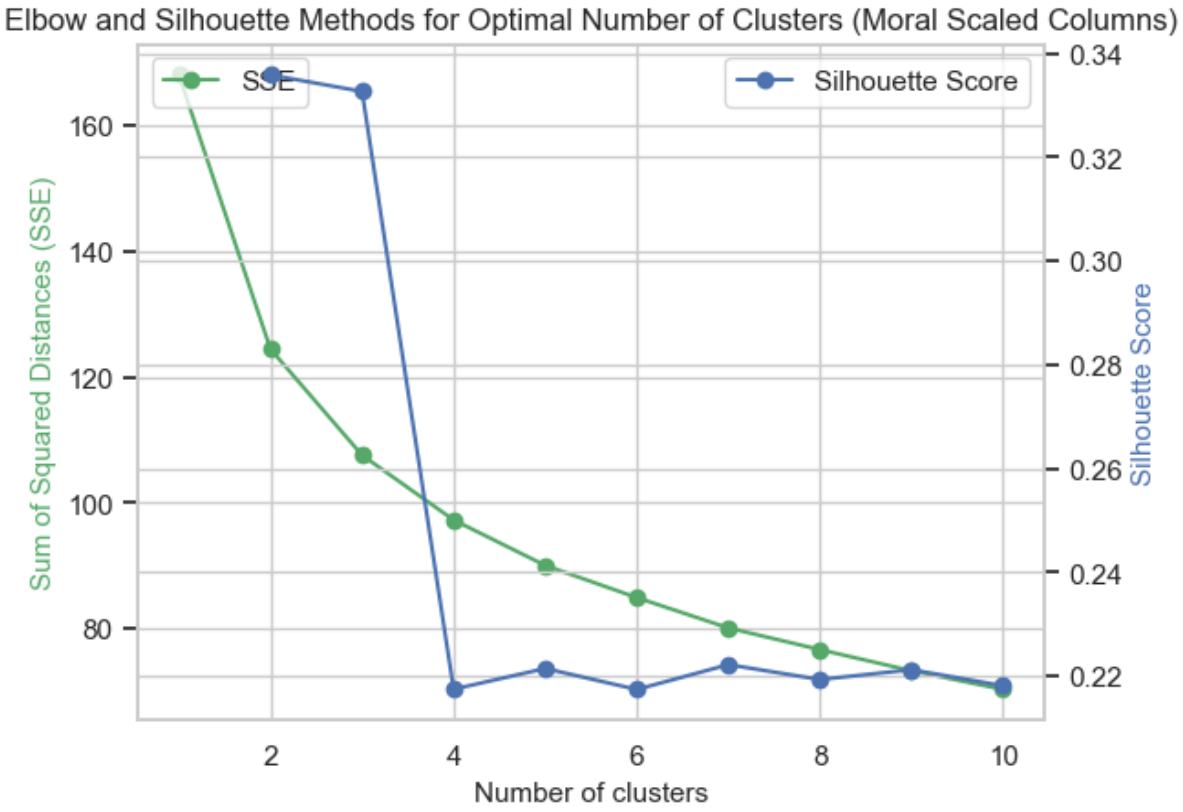

Figure A.17. Elbow and Silhouette Methods with Kmeans Cluster Solution with Moral Scalled Dimensions for Optimal Number of Clusters

Next, we applied K-means clustering based on moral dimensions, resulting in a two-cluster solution. This approach provided a different perspective, showing a clear separation in moral focus between the clusters.

Cluster 0 contained 2,195 articles, focusing predominantly on "General Discussions" (53.35%), "General Politics and Social Issues" (28.52%), "National Politics and Parties" (10.89%), and "Israel-Palestine Conflict" (7.20%). This cluster included 270 authors, representing 75.84% of the total, with top contributors being Miguel Esteves Cardoso (200 articles), Henrique Raposo (100 articles), João Miguel Tavares (63 articles), David Pontes (59 articles), and Afonso Cruz (51 articles).

Cluster 1 contained 1,102 articles, with a more concentrated focus on "General Discussions" (34.39%), "General Politics and Social Issues" (32.12%), "National Politics and Parties" (17.33%), and "Israel-Palestine Conflict" (16.15%). This cluster had 86 authors, representing 24.16% of the total, with top contributors being Carmo Afonso (42 articles), José Miguel Júdice (26 articles), António Barreto (25 articles), São José Almeida (25 articles), and Teresa de Sousa (25 articles).

The moral centroid analysis revealed significant differences between the clusters. Cluster 0 had lower scores across most moral dimensions compared to Cluster 1, with notable

differences in "Authority Virtue" (0.090390 for Cluster 0 vs. 0.276831 for Cluster 1) and "Harm Vice" (0.054939 for Cluster 0 vs. 0.174889 for Cluster 1).

The clustering also effectively separated the authors. For example, Miguel Esteves Cardoso had 97.6% of his articles in Cluster 0 and 2.4% in Cluster 1. In contrast, authors like São José Almeida and Teresa de Sousa had 83.3% and 92.6% of their articles in Cluster 1, respectively, showing a clear distinction in their clustering.

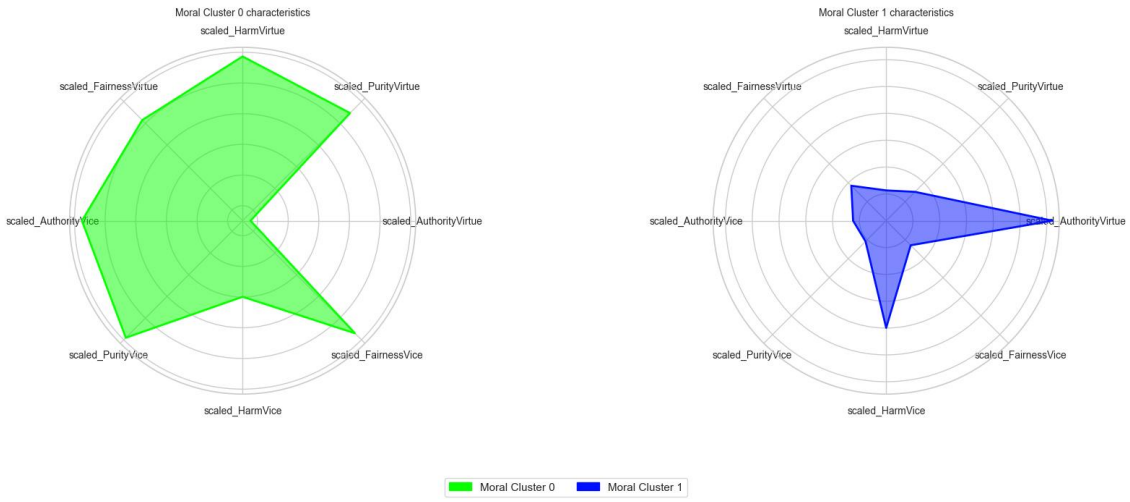

Figure A.18. Two Clusters, Kmeans Cluster Solution, Moral Dimensions profile Radar Chart

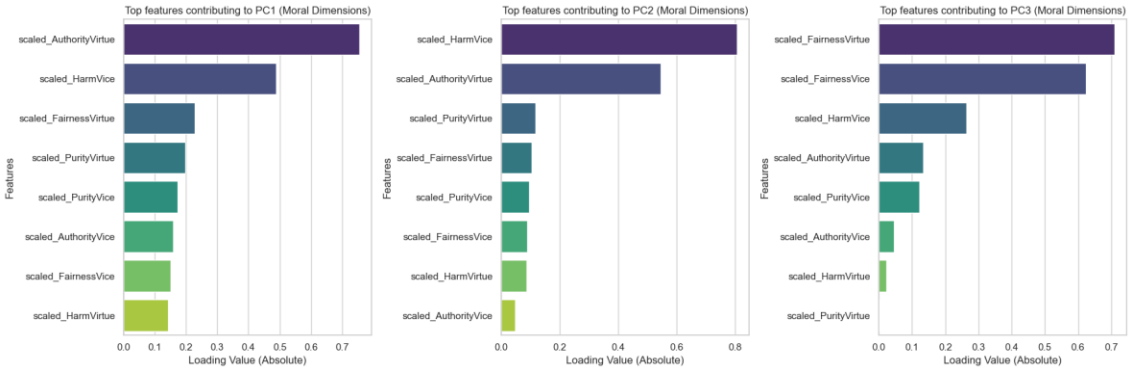

Figure A.19. Two Clusters, Kmeans Cluster Solution Moral Scalled Dimensions, on Top moral dimensions features contribution per PC

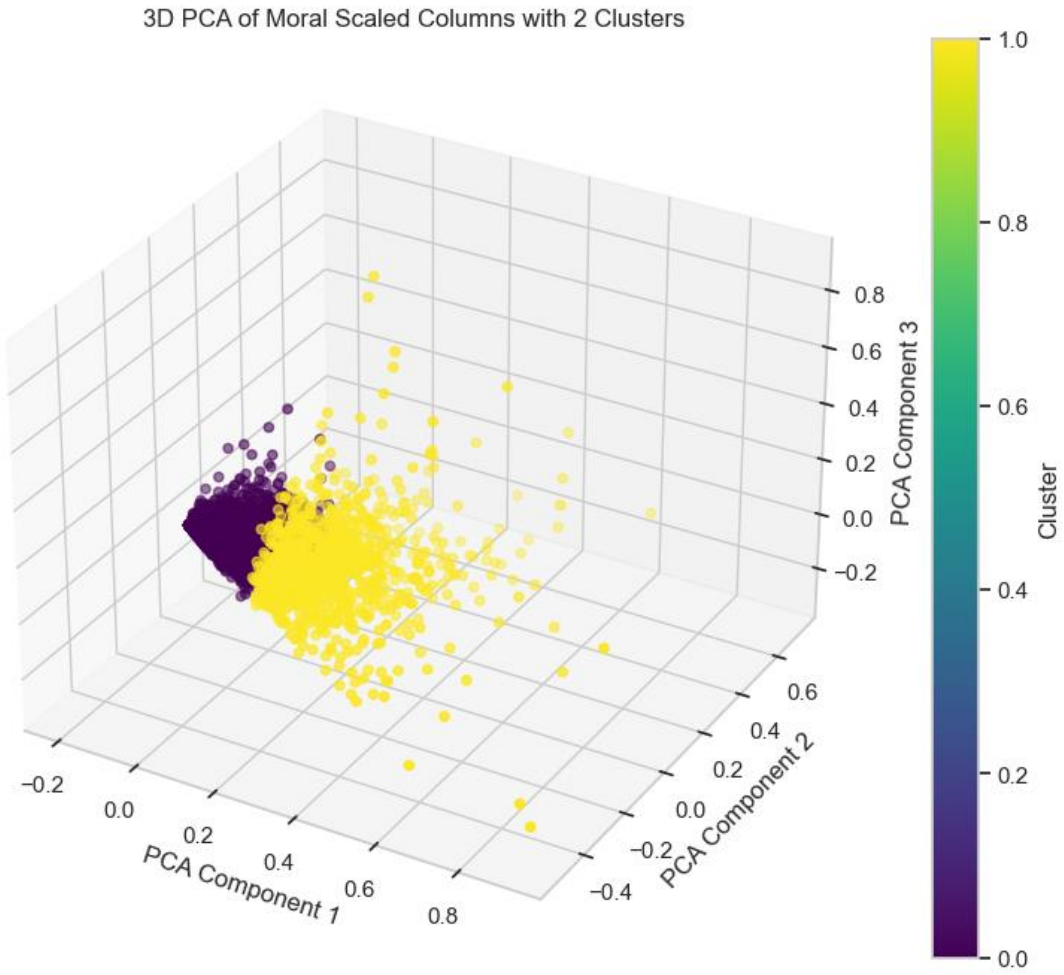

Figure A.20. Cosine Similarity, Kmeans Cluster Solution Moral Scalled Dimensions, 3D Principal Component Analysis plot

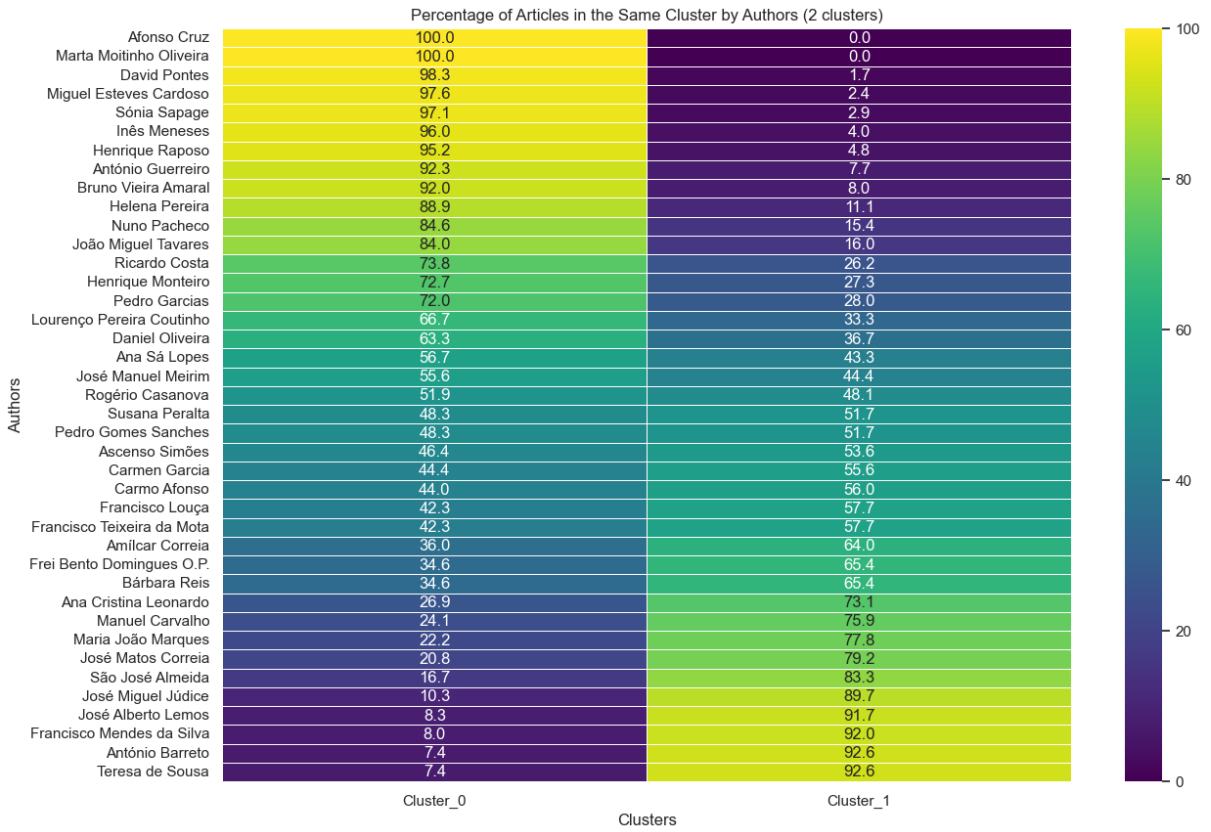

Figure A.21. Kmeans Moral DIM Percentage Matrix of Articles in the Same Cluster by Authors with 2 clusters

Table A.2. Cluster 0 and 1 comparision metrics Two Clusters, Kmeans Cluster Solution Moral Dimensions

Metric	Cluster 0	Cluster 1
Silhouette Score	0.3359	
Davies-Bouldin Index	1.5036	
Number of Articles	2195	1102
Number of Authors	270	86
Percentage of Authors	75.84%	24.16%
Magnitude of Cluster Centroid	0.12	0.36
Percentage of Topics		
General Discussions	53.35%	34.39%
General Politics and Social Issues	28.52%	32.12%
National Politics and Parties	10.89%	17.33%
Israel-Palestine Conflict	7.20%	16.15%
Climate and Energy	0.05%	
Moral Centroids scaled scores		
scaled_AuthorityVirtue	0.09	0.27

Metric	Cluster 0	Cluster 1
scaled_PurityVirtue	0.03	0.07
scaled_HarmVirtue	0.02	0.05
scaled_FairnessVirtue	0.03	0.09
scaled_AuthorityVice	0.02	0.06
scaled_PurityVice	0.02	0.05
scaled_HarmVice	0.05	0.17
scaled_FairnessVice	0.02	0.06

Figure A.22. Two Clusters, Kmeans Cluster Solution, Moral Scalled Dimensions Clusters Characteristics Bar Chart

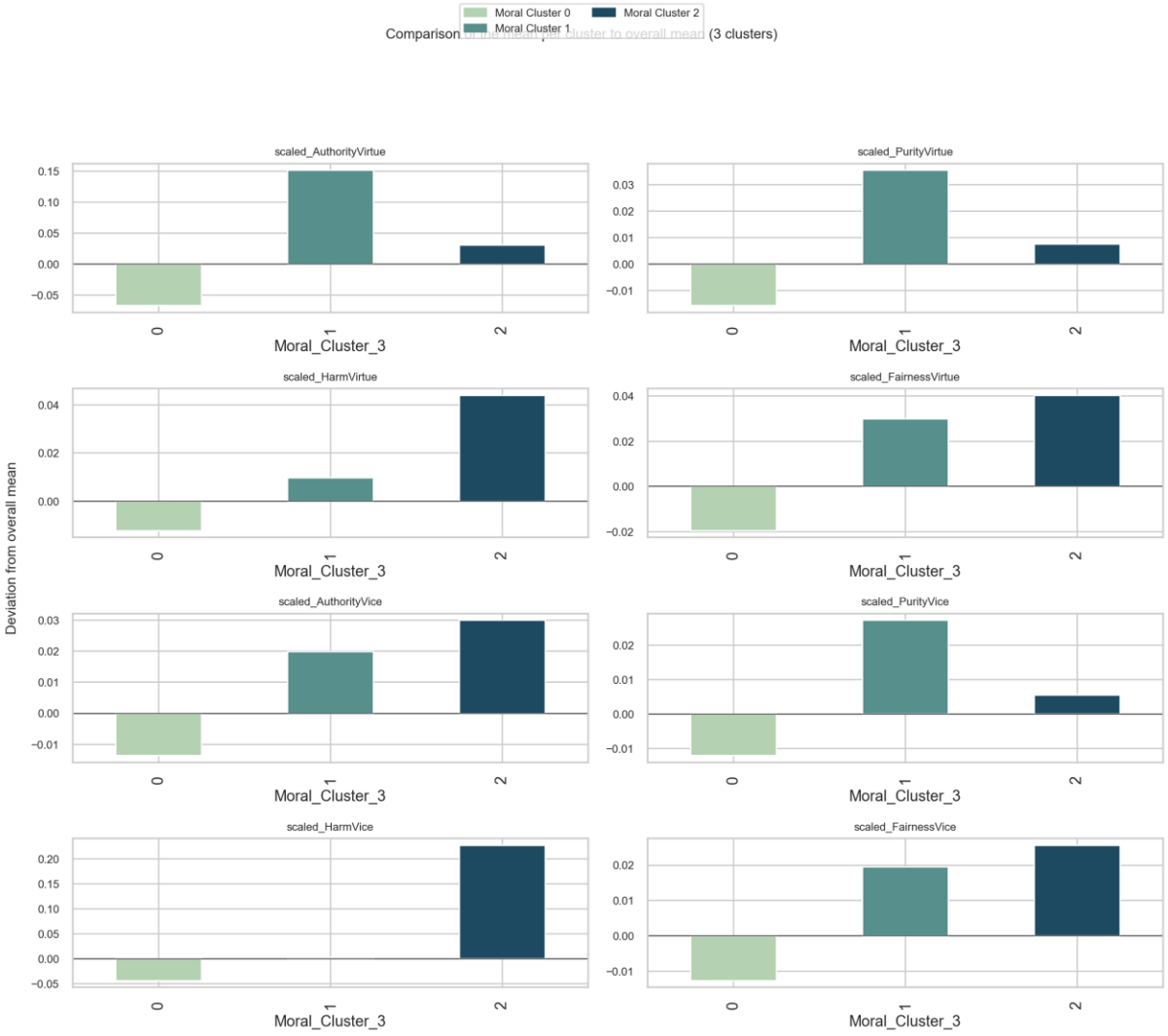

Figure A.23. Three Clusters, Kmeans Cluster Solution, Moral Scalled Dimensions Clusters Characteristics Bar Chart

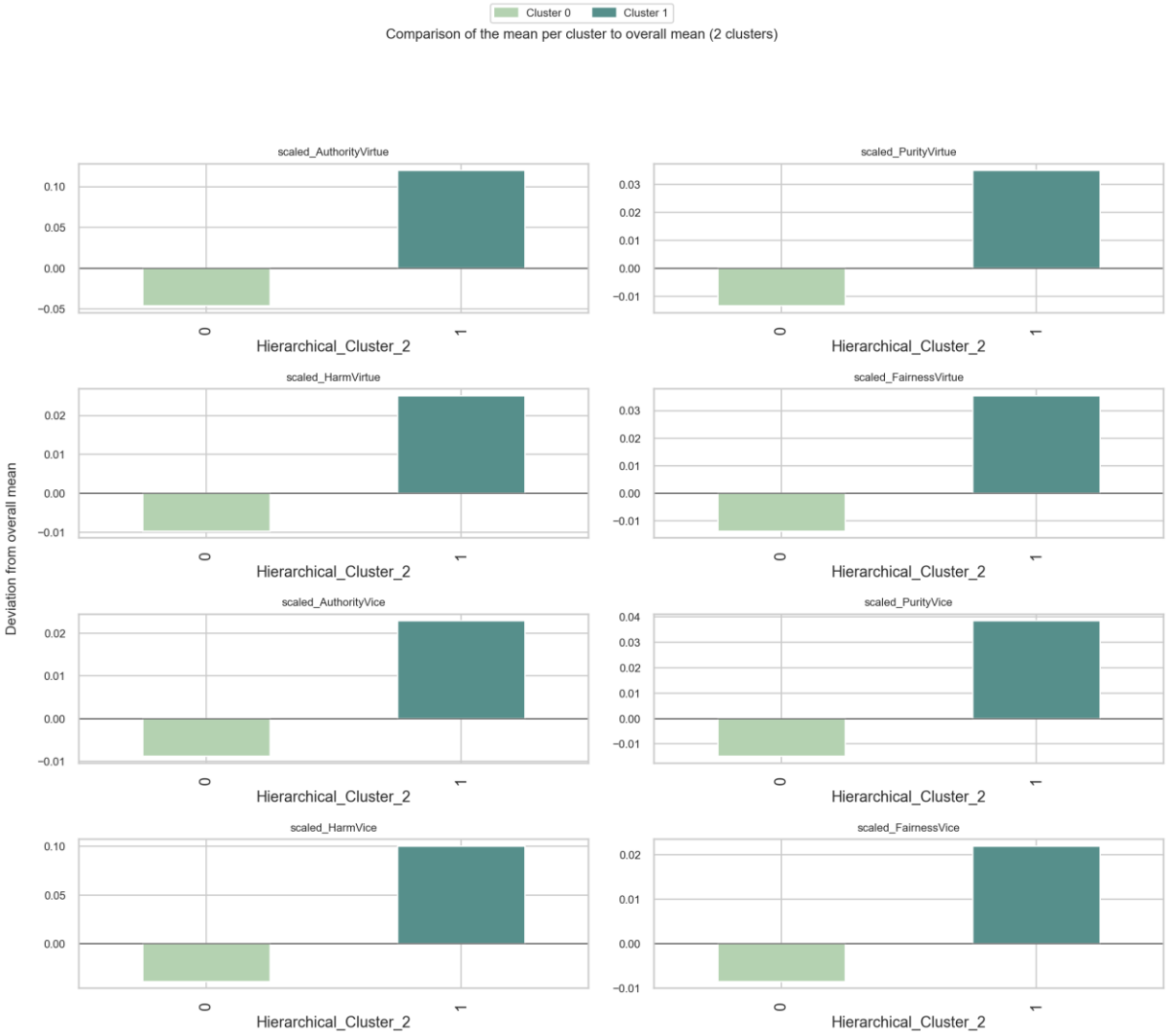

Figure A.24. Two Clusters, Hierarchical Solution, Moral Scalled Dimensions Clusters Characteristics Bar Chart

Figure A.25. Three Clusters, Hierarchical Solution, Moral Scalled Dimensions Clusters Characteristics Bar Chart

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa