Portuguese Far-Right Discourse on Social Media: Insights from Topic Modeling

Mauro Cardoso
Instituto Universitário de Lisboa (ISCTE-IUL), Portugal
INESC-ID Lisboa, Portugal

Eugénio Ribeiro
Instituto Universitário de Lisboa (ISCTE-IUL), Portugal
INESC-ID Lisboa, Portugal

Fernando Batista
Instituto Universitário de Lisboa (ISCTE-IUL), Portugal
INESC-ID Lisboa, Portugal

Abstract

This study analyzes the social media discourse of leading figures from Portugal's far right party CHEGA, examining 10,323 posts on X (formerly Twitter) published between late 2019 and mid-2024. Using BERTopic, 59 latent topics clustered into two main discursive dynamics were found: (1) ideological and public, and (2) party, electoral and parliamentary related. Within the first dynamic, we conducted a focused sub-analysis of themes related with identity, immigration and security narratives – topics that display posting peaks around electoral cycles, suggesting the strategic use of emotionally charged, identitarian frames for political mobilization. The model exhibits strong topic coherence and lexical diversity, indicating its robustness in extracting thematic structures from politically polarized microtexts. Nevertheless, our findings are constrained by source, the absence of interaction metrics, and the unmet need to link online discourse to offline events. This study demonstrates how computational topic modeling can reveal strategic communication patterns in far-right political discourse and underscores the need for cross-platform and interaction-level research to assess broader societal impact.

2012 ACM Subject Classification Computing methodologies → Natural language processing; Human-centered computing → Social network analysis; Computing methodologies → Topic modeling; Computing methodologies → Language resources; Applied computing → Sociology

Keywords and phrases Political Discourse, Topic Modeling, Far-Right, CHEGA (Portugal), Social Media

Digital Object Identifier 10.4230/OASIcs.SLATE.2025.12

Funding This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) (Reference: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020).

1 Introduction

We live in a time when free expression is widely valued, but with that freedom comes a notable increase in polarizing communication practices, particularly on digital platforms. The growth of far-right politics in Portugal reflects a broader trend observed in several European countries in recent years. Factors such as economic crisis, social tensions, and discontent with political elites have created favorable conditions for the emergence of popular disgust and polarized discourse [30, 10].

CHEGA, one of Portugal's most prominent far-right parties, along with its key figures, has capitalized on these tensions, consolidating its presence on the national political scene [14, 21]. The participatory nature of social media reinforces this process, allowing political figures, activists, and supporters to interact, produce content, and share opinions without

14th Symposium on Languages, Applications and Technologies (SLATE 2025).
Editors: Jorge Baptista and José Barateiro; Article No. 12; pp. 12:1–12:16
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the intervention of traditional news outlets [1, 25], often creating echo chambers where their messages are amplified while divergent opinions are marginalized [3]. Through these platforms, far-right actors extend their reach beyond national borders, contributing to a transnational network unified by opposition to globalization and multiculturalism. Digital infrastructures accelerate the circulation of ideologies rooted in fear and resentment, particularly around immigration, national identity, and economic insecurity [16]. Meanwhile, the line between offensive and hateful speech is often unclear, making content moderation and regulation more difficult [5]. Studying these interactions is crucial to understand how individuals and communities navigate discussions on sensitive topics, respond to political and social events, and contribute to the formation of public opinion.

This research aims to explore these online dynamics and their connection to the rise of far-right politics in Portugal, an issue that has raised concerns both nationally and internationally [3]. By analyzing these communication patterns and their underlying structures, this research aims to contribute to a more comprehensive understanding of the digital ecosystem and its impact on contemporary political landscapes. In this context, understanding the topics inherent in this type of data over time is crucial. An established method in Natural Language Processing (NLP) for identifying topics in text is topic modeling. For this research, we used BERTopic [12], a topic-modeling technique that leverages transformer-based sentence embeddings together with a class-based variant of Term Frequency-Inverse Document Frequency (TF-IDF). This robust and flexible approach allows for the capture of semantic and contextual relationships between words, overcoming the limitations of traditional methods such as Latent Dirichlet Allocation (LDA). BERTopic makes it possible to generate more coherent and interpretable thematic representations, as well as supporting dynamic topic modeling, which makes it possible to analyze the temporal evolution of discursive themes in sequential corpora. This functionality is particularly relevant for the study of political discourses and ideological dynamics, as it makes it possible to observe the transformations in topics over time, as demonstrated in the model's applications with Donald Trump's tweets and the transcripts of the United Nations general debates [12]. This research aims to answer the following questions:

- What are the key drivers behind the rise of far-right politics in Portugal and in what ways does the political discourse play a role in shaping and contributing to the growth of these movements?
- What is the role of social media in amplifying the far-right discourse?
- What are the main topics discussed during election periods? How do these topics evolve over time?

This document is organized as follows: Section 2 provides an overview of the related work. Section 3 describes the research methodology, including data collection and processing, as well as model implementation. Section 4 starts with a quantitative evaluation of the models produced and then delves into the analysis of the identified topics and their relationship with the far-right political discourse. Finally, Section 5 summarizes the main findings of the study, discusses its limitations, and provides directions for future research.

2 Related Work

This section introduces the study into the existing literature, beginning by highlighting its social context and then analyzing the most relevant methodological approaches.

2.1 Social Context

The rise of far-right movements in recent decades has generated a substantial interdisciplinary debate. Scholars identify structural and sociocultural drivers, economic pressures, perceived cultural insecurity, and shifting structures of political opportunities while also underscoring how the technical advantages of social media platforms reshape mobilization dynamics. Across national settings, far-right actors rely on a remarkably stable discursive repertoire anchored in nationalism, anti-globalism, anti-elitism, anti-immigration, and cultural nativism. These themes coalesce in narratives that cast society as under attack of external enemies (immigrants or supranational bodies), internal moral decay, and betrayal by domestic elites are presented as simultaneous threats. By positioning themselves as the sole authentic voice of the people, such actors advance xenophobic policy prescriptions while explicitly distancing their programs from fascism and biological racism. Notably, European parties frequently frame the European Union's freedom-of-movement regime as an existential menace, conflating immigration with both economic decline and cultural degradation. Anti-immigration frames therefore remain indispensable for recruiting sectors worried about the growing societal heterogeneity [26].

Until 2019, Portugal stood out in western Europe for the absence of an electorally viable far-right party. The founding of CHEGA ended this phase of exceptionalism, introducing Portuguese politics into the larger European populist wave [20]. The domestic populist field spans multiple dimensions: historical myths, religious invocations, and strategic competition among parties. Media outlets routinely frame populist rhetoric as demagogic or simplistic; strikingly, this label is applied not only to radical actors in CHEGA but also to centrist figures like President Marcelo Rebelo de Sousa whenever emotive, moralizing discourse is employed. Survey data reveal that citizens with populist sympathies resemble habitual abstentionists in institutional distrust, euroscepticism, and perceived identity threat, yet differ markedly in their higher political interest and engagement. These traits strongly predicted support for CHEGA's leader André Ventura in the 2021 presidential election, whereas in the 2022 legislative contest a different determinant, trust in the incumbent government, proved more decisive [20]. Personality analysis also links populist attitudes to conscientiousness, extraversion, neuroticism, and low openness to experience, consistent with a preference for firm immediate solutions in uncertain contexts. Importantly, the Portuguese case shows that populist attitudes can persist without immediate electoral translation, demonstrating the need to consider cultural, institutional, and situational factors together [20].

An online survey carried out between May and June 2021 (n = 3,183) offers the first systematic portrait of CHEGA's militants. Predominantly male, middle-class, educated, and politically active; many without any previous party affiliation, they express great dissatisfaction with democracy, without adopting antidemocratic principles. Instead, they support plebiscitary mechanisms that elevate the will of the people, defending a nativist logic, based on traditional customs, relegating religious identity. Although explicit biological racism is uncommon, essentialist views on culture and ethnicity, especially Roma, are widespread. Aligned with their leader André Ventura, their personalist and punitive populism, characterized by an anti-corruption zeal, calls for strong authority, rejection of multiculturalism, opposition to progressive agendas, emerge as the main drivers of CHEGA's growth, placing the party between protest populism and cultural nationalism [21].

The electoral rise of CHEGA is linked to the confluence of political discourse and weak digital skills among large segments of the Portuguese population. The limited capacity to verify sources exposes citizens to misinformation, while social media algorithms favor sensationalism and polarization of content, reinforcing echo chambers [14]. CHEGA's communication repertoire: fearmongering, stigmatization, anti-elitism, narrative manipulation, achieves high visibility and engagement online. Comparative parallels with Ventura, Bolsonaro, and Trump underscore a shared playbook: manipulation of electoral narratives, strategic use of Artificial Intelligence (AI) in communication, and selective appropriation of democratic symbols that ultimately erode institutional trust. Therefore, scholars recommend media literacy policies, platform regulation, source accountability, and the reinforcement of professional journalism as countermeasures [14]. The offline ramifications are tangible: the mobilization of the digital platform has preceded street protests and, in some cases, incidents of violence [33, 15]. As social networks consolidate as primary platforms for communication [19], researchers now rely on NLP to classify and interpret large volumes of contentious text.

2.2 Computational Methods for Political Discourse Analysis

Through a systematic review of 64 studies [32] on the application of NLP in the analysis of extremism, the field has been mapped in terms of techniques, tools, datasets, and methodological approaches used to describe and detect extremist discourse. The reviewed literature is organized into five analytical dimensions: research topics, techniques employed, empirical applications, available software tools, and accessible datasets. The theoretical framework carefully distinguishes between extremism and radicalization, defining extremism as an antidemocratic ideological movement that may or may not involve violence, while radicalization is conceptualized as a psychological process of detachment from democratic norms. The operational definition of extremist discourse is articulated through five dominant narrative types: political, historical, sociopsychological, instrumental, and theological-moral. These narratives are often accompanied by discursive elements such as hate speech, war metaphors, and dehumanization strategies.

From a technical point of view, the studies reveal a predominance of classical NLP techniques such as TF-IDF, n-grams, and sentiment analysis with feature extraction techniques for syntactic and semantic analysis frequently include Part-of-Speech Tagging (POS), Named Entity Recognition (NER), and topic modeling with LDA. However, there is a growing adoption of deep learning models, showing particular promise in distinguishing extremist discourse, especially when combined with deep learning algorithms. Among the challenges identified are the limited explainability of deep learning models, the complexity of handling multilingual and code-mixed texts, and the lack of publicly available annotated datasets. However, NLP emerges as a robust and rapidly evolving set of tools to detect and characterize extremist narratives, with significant applications in both academic research and the development of public policies and preventive strategies against online radicalization.

Topic modeling has been instrumental in identifying discourse patterns in large volumes of text and has historically been dominated by methods such as LDA that marked a turning point in probabilistic topic modeling for textual corpora by proposing a generative approach grounded in hierarchical Bayesian principles [2]. However, such approaches are based on bag-of-words representations, which limits their ability to capture semantic and contextual relationships between words. When applied to data from social media platforms, the technique must account for the challenges inherent in short, fragmented, and noisy content. LDA remains one of the most widely used probabilistic models for this task; however, its performance is limited in contexts involving microtexts such as tweets due to the sparsity and lack of contextual co-occurrence [18].

Recent studies comparing LDA, Non-negative Matrix Factorization (NMF), Top2Vec, and BERTopic in 31,800 tweets about travel during the COVID-19 pandemic confirm the well-known obstacles of topic modeling in microtexts: brevity, lexical noise, and heterogeneity. The methodological pipeline included standard preprocessing steps (punctuation and stopword removal, lemmatization, and normalization). This type of common preprocessing techniques, including others like textual noise reduction (e.g., eliminating hashtags, URLs, and slang), lower-casing, are critical to ensure data quality. Furthermore, term weighting methods such as TF-IDF are often applied to optimize word relevance and topic representativeness [32, 18]. While LDA tends to generate generic and overlapping topics [8, 18], NMF improves thematic coherence but remains limited to TF-IDF representations. Embedding-based models, in particular BERTopic, overcome these barriers by producing semantically more cohesive, stable, and easily visualizable topics between domains [12], offering modern features that make it scalable and flexible for large, short, or polarized corpora, such as social networks and political discourse studies. However, they are limited to assigning one dominant topic per document and can generate many topics and outliers that require manual inspection.

The study closest to this work, both in terms of subject and methodology, applied BERTopic to a corpus of more than 750,000 tweets about the German federal elections of 2021 [13]. The key topics identified included COVID-19, climate policy, tax policy, digitalization, anti-Semitism, gender equality, and the legalization of cannabis. Although some themes, such as the pandemic and digitalization, were common in all subgroups, others, such as humor and cannabis, appeared more prominently in public mentions. Specific global and national events, such as the Taliban's resurgence in August 2021 and debates surrounding Israel, were reflected in temporary spikes in topic-related activity. Sentiment analysis revealed that discussions around the pandemic and immigration were more frequently associated with negative sentiment, especially in tweets from official accounts. Temporal analysis of topic distributions reveals that certain themes gain prominence during specific time periods, reflecting changes in public attention and discourse engagement. These findings suggest that the integration of topic modeling with temporally-aware and linguistically-informed preprocessing pipelines provides a robust foundation for analyzing discursive trends in political social media data. In the same study, it is noted that although BERTopic was found to be effective in capturing latent themes, some limitations were observed, including challenges in interpreting topics with low coherence and the restriction of limiting the generation of topics, which may have excluded less frequent but significant content. The approach proved effective in extracting complex discursive patterns in electoral contexts, demonstrating the usefulness of BERTopic as a tool for the computational investigation of political communication on digital platforms. The integration of multimodal data, temporal analysis, and sentiment analysis is the key to capture the dynamics and diversity of online political discourse [13].

3 Research Methodology

This section outlines the data collection process, the steps taken to prepare the data, and the setup and implementation of the topic modeling pipeline.

3.1 Data Collection

A total of 10,323 unique posts were extracted using the API of the X platform between late 2019 and mid-2024, focusing on posts from prominent figures in the CHEGA Party, including its leader. The selection of these figures was based on their ranking within the party's official list of members, as published on its website[1], as well as their number of

[1] https://partidochega.pt/index.php/orgaos-nacionais

followers, posting activity, and account longevity. Figure 1 shows the distribution of tweets per source. Furthermore, the party's political performance data by district was manually collected from officials at the Portuguese government site for the same period, with 2019 marking the party's foundation. Metrics were compiled from these data, for example, voter totals over time. The data collected was then transformed, mapped, and stored in tabular structures, facilitating analysis and further exploration. Figure 2 illustrates the temporal distribution of the data, per day. Initially, from 2019 to early 2022, there was a relatively low level of activity, with sporadic spikes in the number of posts. From mid-2022, there is a noticeable increase in the volume of tweets. This trend intensifies significantly throughout 2023 and into 2024, with consistent peaks reaching more than 20 tweets per day, coinciding with the party's electoral growth.

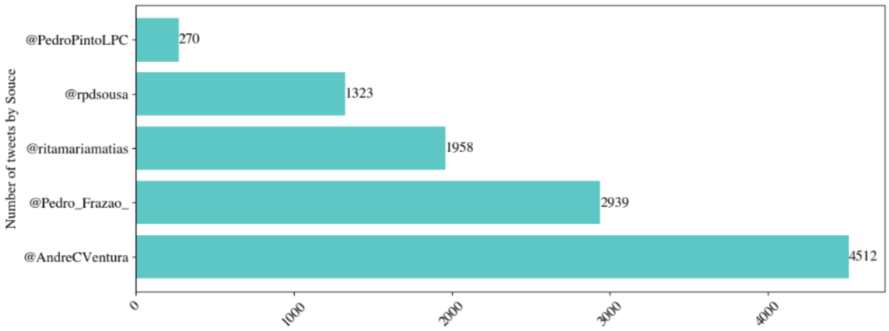

Figure 1 Number of tweets extracted from the account of each of CHEGA's main politicians.

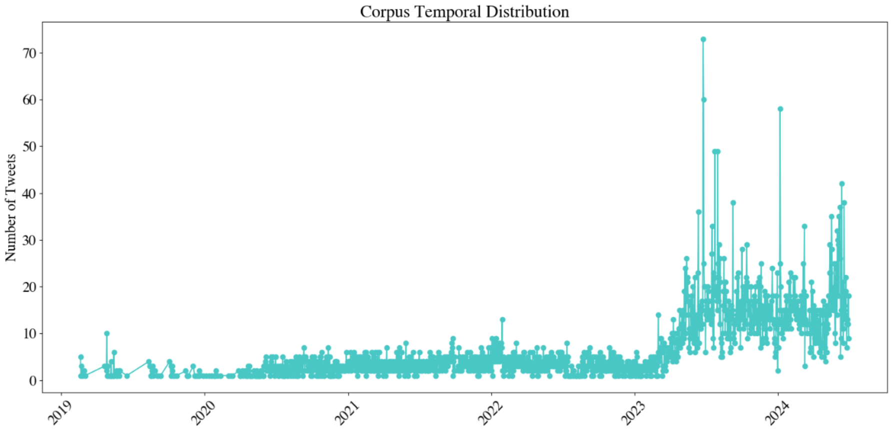

Figure 2 Temporal distribution of the tweets.

3.2 Data Preprocessing

As discussed in Section 2, the preprocessing stage is essential when dealing with social network texts, as informality and network specific mention formats introduce additional problems for analysis and model refinement [17, 32]. The collected texts were subjected to a series of preprocessing steps designed to clean, standardize, and improve their consistency for analysis. Generic mentions of user accounts were removed and/or replaced. The text was then cleaned of URLs, punctuation and special characters, as well as emojis that could interfere with token recognition. Next, the text was tokenized, simultaneously filtering out stopwords that have been aggregated from different sources, to eliminate terms without semantic relevance. After this process, the empty texts were removed. Finally, duplicates (mainly retweets) were removed to avoid redundancies in subsequent tasks.

To ensure consistency of entity representations, a normalization process based on NER techniques was implemented. It arose from the need to consolidate the variants that would appear in the top words between the topics, following the in-depth experimentation described in Section 3.3. Specifically, all variants of the same person or institution were converted into a common identifier. Table 1 shows some examples of this process.

Table 1 Example of substitutions of singular and pair terms.

Original Term	**Substitution**
“andréventura” OR “andreventura” OR “venturas” OR “ventura” OR “andrecventura” OR “andré ventura” OR “andre ventura”	andre_ventura
“partidochega” OR “chega” OR “grupoparlamentarchega” OR “partido chega”	partido_chega
“costa” OR “antonio costa” OR “antónio costa”	antonio_costa
“montenegro”	luis_montenegro
“rocha” OR “rui rocha”	rui_rocha
“cotrim” OR “joão cotrim” OR “joao cotrim”	joao_cotrim
“matias” OR “rita matias”	rita_matias

3.3 Model Implementation

To identify emerging topics, a topic modeling pipeline based on the BERTopic model was implemented, adapted to the linguistic and thematic context of the study. A TF-IDF vectorization technique was applied using 5,000 features, designed to reduce dimensionality and mitigate the impact of sparse terms while preserving a rich representation of textual data. The frequency thresholds were empirically defined, with a minimum document frequency of 1% and a maximum of 85%. These thresholds served two purposes: first, to exclude rare and potentially irrelevant terms that appear in only a small number of documents, thus reducing noise and improving topic stability; and second, to remove overly frequent terms (similar to contextual stopwords) that do not contribute meaningfully to semantic differentiation across documents. This process ensured that the resulting feature space emphasized terms with greater topical relevance and discriminative value.

A customized list of stopwords was incorporated to remove common and politically irrelevant tokens, including elements specific to the social media platform, such as “rt” and “id”. These tokens were identified through multiple iterations of the model output. The pipeline was applied to both the preprocessed data and the raw data.

Text embeddings were generated using the multilingual model Language-agnostic BERT Sentence Embedding (LaBSE) [9], which was trained to capture semantic similarity between languages. We opted for this model as, in preliminary experiments, it led to the generation of more consistent and interpretable topics than the multilingual Sentence BERT models [29] and the Portuguese-specific models of the Serafim family [11].

Dimensionality reduction was performed using Uniform Manifold Approximation and Projection (UMAP) with cosine distance and custom configurations for `random_state=30`, `n_neighbors=5`, and `min_dist=0.03`, ensuring replicability [24]. This step aimed to improve the separation of latent topics in the embedded space. Clustering was performed using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm, configured for increased sensitivity to local point density (`min_cluster_size=50` and `min_samples=3`) [4, 23, 22]. In a context where data density varies, the adjustment and definition of these parameters were experimental to balance granularity and thematic sensitivity. This enabled for more sparse clusters and significant semantic heterogeneity between documents.

3.4 Evaluation Methodology

To perform a solid analysis based on the topics identified by a topic model, it is important to assess the intrinsic quality of the model. For this purpose, we rely on two metrics. Semantic topic coherence is a fundamental step in validating topic models, particularly when the goal is to ensure that the generated topics are interpretable by humans. One of the most widely used metrics for this purpose is c_v coherence, proposed by [31], which measures how well the words of a topic fit semantically by comparing their large-scale co-occurrence patterns with a vector metric, offering an efficient compromise between performance and interpretability. It has proven to be robust even when working with short texts or when topics are generated using modern approaches such as BERTopic. Topic models with higher coherence scores tend to be more semantically interpretable and meaningful. Beyond semantic coherence, topic diversity is another crucial dimension for evaluating topic model quality. Topic diversity quantifies the lexical overlap between topics and is defined as the proportion of unique words across the top-n terms from all generated topics [7, 6]. Values near zero indicate significant topic redundancy, while values approaching one suggest a greater variety and distinction among topics.

4 Result Analysis

This section starts with the quantitative evaluation of the topic model that serves as the basis for our analysis. Then, a combined quantitative and qualitative discussion of the topics is presented, covering their semantic consistency, thematic grouping, and relationships with each other. In line with the related work in Section 2, the findings are expected to revolve predominantly around key themes such as immigration, national identity, corruption, and public safety. These themes are likely to be articulated in a personalized and emotionally charged manner, consistent with populist frames that dramatize a moral divide between virtuous citizens and a corrupt elite. The repeated emphasis on safeguarding order, combating political decadence, and defending nativist values supports a discursive construction of internal or external enemies, thus legitimizing exceptional political interventions centered on strong leadership.

4.1 Model Evaluation

The model identified 59 distinct topics, including the outlier class (topic -1). No application was used to avoid outliers, in order to maintain a more truthful nature of the data [6]. The topics achieved an average coherence score of 0.55, with values ranging from a minimum of 0.29 to a maximum of 0.73. These results indicate a relatively balanced distribution, with most topics displaying satisfactory levels of semantic coherence, particularly considering the nature of the collected data. Regarding topic diversity, the results indicate a high level of lexical diversity between topics, with a maximum value of 0.96 when considering the top 10 terms in each topic. Even with an increase in the number of words analyzed per topic, diversity remains high, 0.92 (top-20) and 0.88 (top-30), showing that the repetition of terms between different topics is low. The slight decrease in diversity as the number of terms evaluated increases is expected, given the increased likelihood of lexical overlap. However, the value of 0.81 for the top-50 terms continues to indicate a good thematic separation between topics, reinforcing the robustness of the model in capturing the distinct dimensions of the discourse.

4.2 Topic Analysis

A short description was attributed to each topic by providing its top 50 keywords and the domain context to the GPT-4o Large Language Model (LLM) [28, 27]. The generated descriptions were then manually curated. Considering the hierarchical nature of the topics generated by BERTopic, this process also enabled the preliminary grouping of topics into thematic categories, as shown in Table 2 and Figure 3.

Together, the groups map ten broad domains of the party's discourse: institutional politics and electoral strategy; corruption and scandals; ideological conflict and online polarization; national identity and historical memory; European and transnational far-right coordination; social tensions around migration, crime and minority rights; religion and moral issues; economic hardship and public health crises; media and communication dynamics; and a residual class that gathers low-frequency or heterogeneous items.

Figure 3 represents the thematic hierarchy map, together with the corresponding topic identifiers (T) and the number of associated documents (D). The groups were supported and explored with hierarchical clustering (as provided by BERTopic). This structure supported the assignment of broader interpretative meanings, resulting in a hierarchically organized thematic framework. Three overarching domains emerged: (0) Other or outliers (2,382; 23%), being a residual class; (1) Ideological and public dynamics (2 981; 29%), divided between security-identity discourses (1 017 documents) and moral-religious discourses (797), accompanied by 1 037 colloquial digital speeches that amplify these frames; and (2) Party, electoral, and parliamentary dynamics (4,960; 48%), covered by 1,094 texts on public policies, leadership and economy, 1,007 on territorial mobilization, and above all, 2,859 disputative-style speeches combining accusations of corruption with campaign communication, highlighting the party's central media strategy. Taken together, the hierarchy exposes three interdependent rhetorical pillars: an identitarian securitarianism based on the "us" versus "them" dichotomy, a countercultural moralism oriented toward defending traditional values, and a dramatized antisystem denunciation, disseminated through strong media and digital presence. The uneven distribution of the documents, with almost half dedicated to political confrontation and around a third to the identity-moral axis, suggests that the visibility gained in parliamentary debates and in the media serves as a vehicle for reinforcing emotive messages of security, nationalism, and moral order. This empirical framework thus offers a robust starting point for temporal analyses of the party's thematic evolution and for checking how peaks in public attention correlate with the strategic use of certain frames.

Figure 4 shows the aggregate monthly evolution of the three main thematic dynamics mentioned previously, from late 2019 to mid-2024, highlighting the relevant electoral moments marked in red. The graph shows not only the volume of monthly mentions within each category, but also allows us to observe patterns of discursive intensification associated with key moments in the national political calendar, such as legislative and presidential elections. The overlap of curves and markers indicates a possible correlation between the increase in discursive activity on X (formerly Twitter) and electoral cycles, reflecting both the growth of the party's support base and its strategic positioning on social networks over time.

1. Others: – **T:** -1 **D:** 2382

2. Ideological and Public Dynamics

- **2.1** Public Order, Clash of Values, Religious and Identitarian Issues
  - **2.1.1** Conservative Approach, Public Order, Nationalism/Patriotism and Traditional Values
    - **2.1.1.1** Social, Political Accountability and Dissatisfaction – **T:** 45, 47 **D:** 130
    - **2.1.1.2** National Identity and Tradition, Ethnic Issues, Immigration and Security – **T:** 7, 12, 14, 21, 39, 41, 43, 48 **D:** 1017
  - **2.1.2** Ideological Positions, Moral, Sexual and Religious Debates
    - **2.1.2.1** Sexual Crimes, Feminism/Abortion, Religious Education and Gender/Sexuality Debates – **T:** 24, 27, 44, 46 **D:** 357
    - **2.1.2.2** Religious Conflicts, Political Ideologies, Extreme Disputes, Relations between Belief, Morals and Politics – **T:** 8, 32, 34, 50 **D:** 440
- **2.2** Discourse of Social and Digital Interaction, Internet Colloquial Tone – **T:** 5, 6, 29, 35, 37, 38, 40, 57 **D:** 1037

3. Party, Electoral and Parliamentary Dynamics

- **3.1** Governance, Economy and Social Policies, Public Health
  - **3.1.1** Public Management, Health, Investigations/Scandals and Power Conflicts – **T:** 22, 30, 31, 33 **D:** 390
  - **3.1.2** Socio-Economic Criticism and Debate – **T:** 11, 16, 23, 51 **D:** 539
  - **3.1.3** Controversies and Disputes about Political Leadership and Government Management – **T:** 53, 56, 58 **D:** 165
- **3.2** Political-Electoral Dynamics, Public Management, Ideological Debates and Media Coverage
  - **3.2.1** Political Dispute, Ideological Positions, Communication and Media – **T:** 0, 1, 2, 3, 4, 10, 15, 18, 25, 26, 36, 42, 52, 54, 55 **D:** 2859
  - **3.2.2** Political-Parliamentary and Electoral Cycle, Mobilization, Interaction and Event Organization – **T:** 9, 13, 17, 19, 20, 28, 49 **D:** 1007

Figure 3 Thematic hierarchy map.

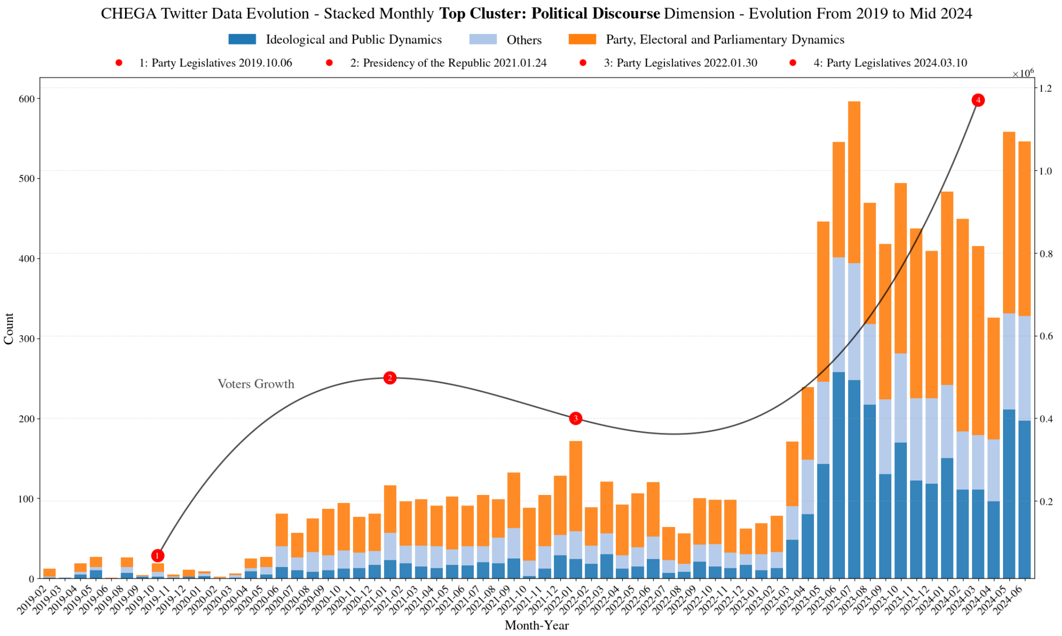

Figure 4 Monthly Evolution of Political Discourse on X (formerly Twitter) associated with CHEGA (2019-2024).

Table 2 Topic-model labels assigned to the 59 BERTopic clusters.

ID	Definition
-1	Others
0	Communication, the Press and Political Mobilization
1	Political Vandalism and Ideological Conflict
2	Corruption and Abuse of Power
3	Political Coalitions, Controversies and Political Responsibility
4	Media Events and Political Events and Campaigning
5	Informal Expressions and Personal Interaction
6	Online Debates and Interpersonal Criticism
7	Social Tensions and Immigration
8	Indoctrination and Ideological Polarization
9	Parliamentary Activities and Constitutional Review
10	Elections and Political Strategy
11	Tax Burden and Social Costs
12	Crime and Public Safety
13	Political Events and Campaigning
14	Immigration and Border Control
15	European Right and Political Coordination
16	Economic and Social Crisis
17	National Identity and Patriotism, Youth and Education
18	Political and Institutional Conflicts
19	Agriculture, Fisheries and Rural Development
20	Local and Regional Elections
21	Family, Faith, Celebrations and Religion
22	Crisis and Collapse in Health
23	Corruption, Public Management and Clientelism
24	Criminal Justice, Violence and Crime
25	Alternative, Criticism and Political Confidence
26	European Right and National Identity
27	Gender Equality and Feminist Dynamics
28	Elections, Regional Campaigns and International Far-Right Movements
29	Insults and Ideological Conflicts, Disinformation and Public Exposure

ID	Definition
30	Diplomacy, Budget and Government Relations
31	Governance, Controversies and Political Responsibility
32	Patriotism, Controversy and Mobilization
33	Opposition, Criticism and Political Management
34	Ideologies and Individual Freedoms
35	Social Interactions, Congratulations and Emotions
36	Controversies and Political Scandals
37	Expressions, Rhetoric and Emotions
38	Racism, Public Opinion and Prejudice
39	Public Safety, Crime and Minorities
40	Disinformation, Defamation and Social Networks
41	Colonialism, Historical Memory and Reparations
42	European Politics, Territorial Action and Institutional Participation
43	Patriotism, National Identity and Historical Memory
44	Indoctrination, Religion and Moral Education
45	Protest, Indignation and Political Outcry
46	Sexuality, Ideology and Child Protection
47	Corruption, Impunity and Criticism of the Political System
48	National Honor, State Careers and Patriotic Duty
49	Public Space, Mobility and Local Identity
50	Religious Conflicts and Terrorism
51	Justice, Cost of Living and Economy, Social Exclusion
52	Elections and Political Participation
53	Government Crisis and Public Administration
54	Political Participation and Institutional Action
55	Political Mobilization and Campaign Rhetoric
56	Public Health and the Hospital System
57	Media, Journalism and Television Representation
58	Participation and Political Figures

For analytical purposes, we have selected the category "national identity and tradition, ethnic issues, immigration, and security", which is a predominant area of focus within this study. It comprises 1,017 documents, representing approximately 10% of the dataset, and belongs to the category of ideological and public dynamics. Table 3 shows eight topics from the previously selected category. Each topic is associated with a meaningful label and is represented by the top-50 most relevant keywords of the topic.

The three topics directly related to immigration and public order (topics 7, 12 and 14) share the words "threat", "violence", and "lack of control", which reinforces the idea that internal security and border management merge into a single narrative about external risk. In contrast, topic 21 ("Family, Faith, Celebrations, and Religion") shifts the focus from security anxiety to a register of values and affective belonging. Words like "brotherhood', "Fátima", "fatherhood", and "happyfamily" refer to Catholic festivities, family cohesion, and a positive imaginary community. This vocabulary acts as a discursive counterbalance: Where the migratory topics point to a threat, this one evokes the protection of traditions. Topics 39, 41, 43 and 48 fall between these two poles. Topic 39 takes up the "crime" but adds ethnic markers ("gypsies", "negro"), indicating that public security is also framed by specific minority categories. Topics 41 and 43 invoke the historical past ("carnation", "Salazar", "Camões") to legitimize a selective memory of the nation, sometimes celebratory ("heroes", "homeland", "pride") and sometimes resentful ("reparations", "traitors"). Finally, topic 48 converges on the valorization of state careers ("military", "firefighters", "police officers") and the "homeland" as a duty, connecting national honor with corporate recognition.

Table 3 Selected topics, manually categorized, representing the national identity, tradition, ethnic issues, immigration, and security.

ID	Topic Definition	Top 50 Keywords
7	Social Tensions and Immigration	veremos, confiar, martim, moniz, oferecer, envergonham, atlântico, ex, seleção, limpo, limpeza, nepalês, louvor, paris, lembrome, policial, juntar, acordar, suíça, convosco, terminamos, amadora, ficaria, contamos, prontos, resistir, podridão, traficantes, mostram, hino, comunidade, destino, revolta, segredo, falava, ferro, sai, bocado, indivíduos, merecem, fraudes, frança, anedota, ocasião, abortar, marine, básicos, pais, islamização, pedras, rainha, símbolo
12	Crime and Public Safety	terror, faca, assaltos, adulto, tiroteio, esfaqueou, multiculturalismo, imigrante, moradores, indivíduo, islâmico, permitimos, incêndio, chamas, rixa, afegão, provocam, agressões, feridos, policial, insegurança, destruídos, armado, incendiários, agredir, orações, ourém, pornografia, plena, drogas, consumo, atirados, martim, moniz, terrorista, fumar, acontecem, ferida, islamização, metro, homicídio, brutalmente, violentos, cenário, agressor, bárbara, agente, infantil, ignoram, droga, psp
14	Immigration and Border Control	lampedusa, ilegais, deportação, aima, imigrantes, integração, imigrante, deportar, migração, massiva, tendas, quotas, refugiados, agência, descontrolados, espadas, asilo, facas, desembarque, refugiado, gratuito, migrações, migrantes, habitantes, apresentarem, dupla, entram, seguras, pacto, indiano, legalização, pedidos, controlada, utilizam, descontrolada, descontrolo, afegãos, idade, regras, chegada, fronteiras, fiscalização, zonas, alterar, chegam, ilegal, nacionalidade, chegaram, recorde, encontram
21	Family, Faith, Celebrations and Religion	brotherhood, tabuleiros, familyparty, happyfamily, fatherhood, familianumerosa, brothers, amam, catedral, parentalidade, family, santo, califórnia, festas, ventre, toiro, espírito, áfrica, irmãos, bispo, salomé, espada, mães, alegria, jesus, magnífica, colegas, natal, jerusalém, estátua, humildes, fátima, companheiros, international, nascimento, santidade, desempenha, paixão, reta, partidária, santuário, activista, dedicação, henriques, palestra, muçulmanos, emocional, católico, empresarial, cartaxo
39	Public Safety, Crime and Minorities	ciganos, magistrados, policiais, criminalidade, elogiar, gnr, seguro, monsaraz, negro, cigano, estatísticas, agir, probabilidade, crimes, violentos, reguengos, polícia, polícias, ligações, criminosos, bandidos, autoridade, publicações, luxo, agressores, agressões, mortos, diariamente, resistiremos, delinquentes, referente, trata, nacionalidade, acontecem, seixal, serem, culpas, coitadinho, homicídio, cigana, colonial, escravatura, prisionais, morto, erro, bandido, nulo, forças, arriscar, perderem, meios
41	Colonialism, Historical Memory and Reparations	cartoon, excolónias, orgulha, rejeitar, traidores, lgbti, desculpa, domínio, mansão, territórios, obscuros, humilhar, vergonhosa, antigas, degradação, antepassados, indemnizações, armadas, impostos, moçambique, desrespeito, forças, governa, pagar, promove, contentes, manter, colapso, corruptos, segurança, tratados, ações, pedimos, protege, pediu, ana_gomes, agem, insultar, compensações, online, cultural, colonial, paga, ilegal, imóveis, sintra, garantir, humanos, prec, polícia, fatura
43	Patriotism, National Identity and Historical Memory	milenar, históricos, lutam, tradições, agente, cravos, camões, dou, colectiva, antepassados, carneiro, perdida, amamos, psp, necessárias, desculpa, brutalmente, sá, portugalidade, símbolos, comunidades, palhaçada, lutaram, identidade, medalha, agredido, espírito, feitos, pátria, orgulho, portuguesas, heróis, hino, fantasma, espadas, protege, serviu, londres, veste, continuarmos, endémica, metro, racismo, independência, escumalha, fábio, deixarmos, bonitos, pergunta, salazar
48	National Honor, State Careers and Patriotic Duty	feriado, reivindicações, queridos, divulgação, policiais, militares, justa, inspiração, novembro, homenagem, entidades, exército, militar, agradecimento, prometido, heróis, sacrifício, protesto, lutaram, carreiras, combatentes, forças, lealdade, data, altas, bombeiros, antigos, empenhados, segurança, corpo, isabel, celebrar, esquecidos, básicos, ideologias, totalitarismo, membro, rainha, prisionais, castelo, seguido, delito, particularmente, operacional, proteção, humana, pátria, armadas, testemunhar, patriotismo

Figure 5 shows the monthly evolution of topics included in the dimension of national identity, tradition, ethnic issues, immigration, and security in the political discourse associated with CHEGA on X. The distribution shows a progressive increase in the incidence of topics related to immigration, public safety, crime, nationalism, and social tensions. The graph shows not only the volume of monthly mentions within each category, but also allows us to observe patterns of discursive intensification associated with key moments in the national political calendar, such as legislative and presidential elections. The overlap of curves and markers indicates a possible correlation between the increase in discursive activity on X and electoral cycles, reflecting both the growth of the support base of the party and its strategic positioning on social networks over time. The pattern of discursive activity remains consistent with that shown in Figure 4. The relationship between the curve and the electoral

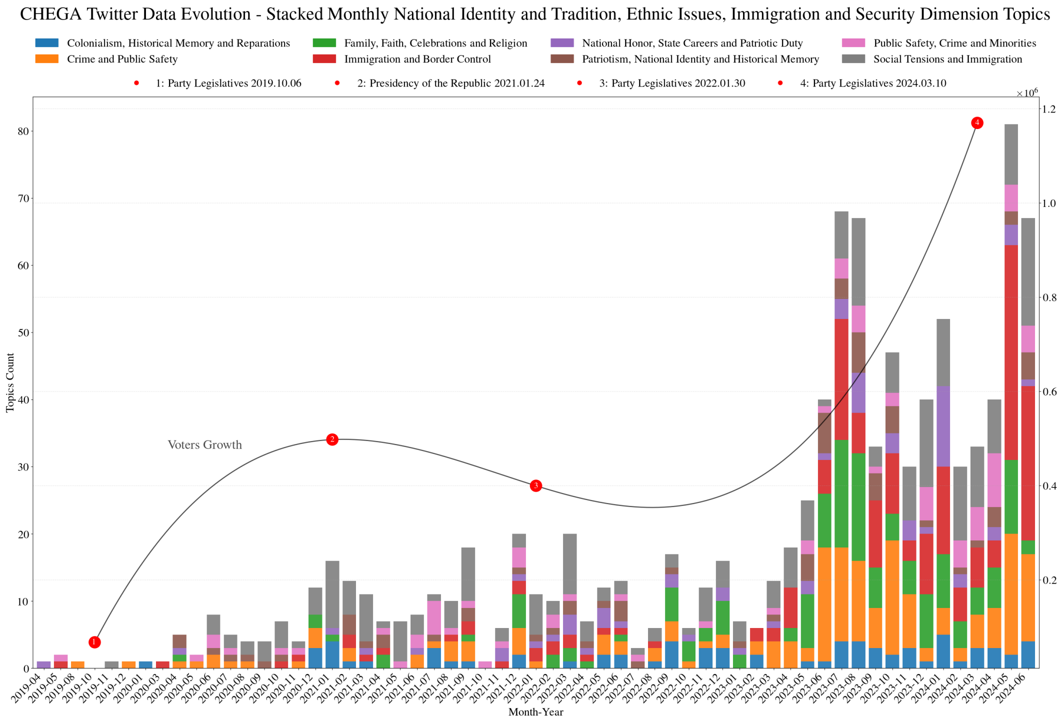

Figure 5 Monthly Evolution of Topics Related to National Identity, Tradition, Ethnic Issues, Immigration, and Security Dimensions (2019-2024).

cycles is also maintained. This growth becomes particularly expressive throughout 2023 and into the early months of 2024, culminating in a sharp discursive peak just before and after the legislative elections in March 2024. This is reflected in the growth in voting numbers, with the range of figures in the 2022 parliamentary elections being around 400,000, rising to more than a million in 2024.

Among the most recurrent subthemes are "social tensions and immigration", "crime and public safety", "immigration and border control", and "family, faith, celebrations, and religion". In particular, social tensions and immigration appear more consistently throughout the timeline, suggesting its centrality in the long-term discursive strategy of the party. Collectively, these themes account for a substantial portion of the discursive volume within this dimension, indicating a communication strategy grounded in the mobilization of negative affect, particularly fear, insecurity, and perceived threat, often linked to the presence of minorities or immigrants, and emotionally reinforced through references to family, faith, and religious values. Further strengthening this divisive narrative, topics such as "colonialism, historical memory, and reparations", "patriotism, national identity, and historical memory", and "national honor, state careers, and patriotic duty" also appear consistently.

These topics suggest a symbolic articulation between security, identity, and morality, constructing a vision of national cohesion under threat. The temporal overlap between the rise in discursive activity and the electoral cycles indicates an instrumental use of these narratives for political mobilization. The overlaid curve, representing the electoral growth of the party, appears to align with the intensification of the security and identity-oriented rhetoric, strengthening the hypothesis that these themes play a strategic role in consolidating and expanding the voter base of the party.

5 Conclusions and Future Work

This study analyzed 10,323 unique posts on X (formerly Twitter) published by key figures in CHEGA between late 2019 and mid-2024, using BERTopic to uncover 59 latent topics in the political discourse of the party. The evaluation of the model revealed an average topic coherence of 0.55 and a topic diversity exceeding 0.80 of the top 50 terms, validating the robustness of the approach. In terms of empirical contributions, hierarchical aggregation enabled the identification of two major discursive dynamics: (1) ideological and public, and (2) political, electoral, and parliamentary. Within the first dynamic, the national identity, tradition, immigration and security subcategory, which represents approximately 10% of the corpus, followed the general pattern of discursive growth observed across the dataset, particularly during electoral periods, culminating in between early 2023 and mid 2024. Although this trend reflects the political nature of the corpus as a whole, the prominence of this subcategory suggests a communication strategy that takes advantage of negative emotional appeals (fear, insecurity), symbolically reinforced through identity and moral narratives, to amplify the political message.

The combination of LaBSE embeddings, UMAP-based dimensionality reduction, and HDBSCAN clustering, together with a recent instruction-tuned LLM (GPT-4o [28, 27]) and manual curation, proved effective in analyzing multilingual and politically polarized corpora. The convergence between peaks in security and identity-related discourse and electoral cycles underscores the need for media literacy policies and content moderation strategies that discourage the dissemination of messages exploiting fear or aversion to the *other*.

However, it is important to acknowledge the potential biases in this study. On the one hand, relying on a single data source introduces selection bias: the focus on the accounts of CHEGA key figures excludes interactions with supporters and counter-discourses. Additionally, the use of monthly snapshots limits the temporal resolution of the analysis, preventing a more fine-grained estimation of causal relationships between offline events and topic fluctuations. On the other hand, although we attempted to minimize human-introduced bias in the topic generation process (e.g. by supporting decisions with quantitative metrics such as c_v coherence and lexical diversity, and by using an LLM to generate the short descriptions of the topics) the manual curation and interpretation of topics remains subjective and susceptible to bias.

Future work should focus on expanding the sample and triangulating across platforms by incorporating data from additional social networks to assess narrative consistency and detect cross-platform variation. Including replies, for instance, would help map the interactional ecology and organic reach of the messages. Estimating causal links between discursive peaks and real-world events remains a critical objective for understanding the dynamics at play.

In the BERTopic results derived from hierarchical clustering, a notably high proportion of cases (23%) were considered outliers (represented by “Others”). While this category is often overlooked, its significant size warrants further analysis and interpretation.

Finally, developing an interactive and comprehensive artifact that includes the full topic analysis can increase the study’s accessibility and public relevance. It can also support more detailed classifications, such as identifying emotional tones or rhetorical strategies within thematic dimensions and associated topics.

References

1 Rémi Almodt. The Right-Wing Perspective: Populist Frames and Agenda on Facebook in Central and Eastern Europe. *Central European Journal of Communication (CEJC)*, 15(3(32)):434–463, 2023. doi:10.51480/1899-5101.15.3(32).6.

2 David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003. doi:10.5555/944919.944937.

3 Manuela Caiani and Patricia Kröll. The Transnationalization of the Extreme Right and the Use of the Internet. *International Journal of Comparative and Applied Criminal Justice*, 39(4):331–351, 2014. doi:10.1080/01924036.2014.973050.

4 Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-Based Clustering Based on Hierarchical Density Estimates. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 160–172, 2013. doi:10.1007/978-3-642-37456-2_14.

5 Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, pages 512–515, 2017. doi:10.1609/icwsm.v11i1.14955.

6 Muriël de Groot, Mohammad Aliannejadi, and Marcel R. Haas. Experiments on Generalizability of BERTopic on Multi-Domain Short Text. In *Widening Natural Language Processing (WiNLP)*, 2022. doi:10.48550/arXiv.2212.08459.

7 Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020. doi:10.1162/tacl_a_00325.

8 Roman Egger and Joanne Yu. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7:886498, 2022. doi:10.3389/fsoc.2022.886498.

9 Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT Sentence Embedding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1: Long Papers, pages 878–891, 2022. doi:10.18653/v1/2022.acl-long.62.

10 Matt Golder. Explaining Variation in the Success of Extreme Right Parties in Western Europe. *Comparative Political Studies*, 36(4):432–466, 2003. doi:10.1177/0010414003251176.

11 Luís Gomes, António Branco, João Silva, João Rodrigues, and Rodrigo Santos. Open Sentence Embeddings for Portuguese with the Serafim PT* Encoders Family. In *Proceedings of the EPIA Conference on Artificial Intelligence*, pages 267–279, 2024. doi:10.1007/978-3-031-73503-5_22.

12 Maarten Grootendorst. BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure. *Computing Research Repository*, arXiv:2203.05794, 2022. doi:10.48550/arXiv.2203.05794.

13 Nils Constantin Hellwig, Jakob Fehle, Markus Bink, Thomas Schmidt, and Christian Wolff. Exploring Twitter Discourse with BERTopic: Topic Modeling of Tweets Related to the Major German Parties during the 2021 German Federal Election. *International Journal of Speech Technology*, 27:901–921, 2024. doi:10.1007/s10772-024-10142-4.

14 Joana Jerónimo. Comunicação na Era da Desinformação: o Crescimento da Extrema-Direita e a Iliteracia Digital. *The Trends Hub*, 1(4), 2024. doi:10.34630/tth.vi4.5683.

15 Ofra Klein and Jasper Muis. Online Discontent: Comparing Western European Far-Right Groups on Facebook. *European Societies*, 21(4):540–562, 2019. doi:10.1080/14616696.2018.1494293.

16 Tiago Lapa and Branco di Fátima. Hate Speech Among Security Forces in Portugal. *Communication Library*, 7:277–293, 2023. doi:10.25768/654-916-9.

17 Xiaohua Liu, Ming Zhou, Xiangyang Zhou, Zhongyang Fu, and Furu Wei. Joint Inference of Named Entity Recognition and Normalization for Tweets. In *Proceedings of the Annual*

Meeting of the Association for Computational Linguistics (ACL), volume 1: Long Papers, pages 526–535, 2012. URL: https://aclanthology.org/P12-1055/.

18 Diana Lopes-Teixeira, Fernando Batista, and Ricardo Ribeiro. Discovering Trends in Brand Interest through Topic Models. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, pages 245–252, 2018. doi:10.5220/0006936202450252.

19 Adrian Lüders, Alejandro Dinkelberg, and Michael Quayle. Becoming "us" in Digital Spaces: How Online Users Creatively and Strategically Exploit Social Media Affordances to Build Up Social Identity. *Acta Psychologica*, 228, 2022. doi:10.1016/j.actpsy.2022.103643.

20 Luca Manucci. Portuguese Populism: People, Parties, and Politics. *Análise Social*, 59(251), 2024. doi:10.31447/202200.

21 Riccardo Marchi and José Pedro Zúquete. Far Right Populism in Portugal: The Political Culture of Chega's Members. *Análise Social*, 59(251), 2024. doi:10.31447/2022116.

22 Leland McInnes and John Healy. Accelerated Hierarchical Density Clustering. In *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 33–42, 2017. doi:10.1109/ICDMW.2017.12.

23 Leland McInnes, John Healy, and Steve Astels. HDBSCAN: Hierarchical Density Based Clustering. *The Journal of Open Source Software*, 2(11):205, 2017. doi:10.21105/joss.00205.

24 Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Computing Research Repository*, arXiv:1802.03426, 2018. doi:10.48550/arXiv.1802.03426.

25 George Newth. 'Talking About' the Far Right and Common Sense. A Case Study of Matteo Salvini's Buon Senso Trope on Twitter (2018–2023). *Acta Politica*, 60:361–384, 2025. doi:10.1057/s41269-023-00327-1.

26 Seyed Nader Nourbakhsh, Seyyed Abbas Ahmadi, Qiuomars Yazdanpanah Dero, and Abdolreza Faraji Rad. Rise of the Far Right Parties in Europe: from Nationalism to Euroscepticism. *Geopolitics Quarterly*, 18(4):47–70, 2021. URL: https://journal.iag.ir/article_129481.html.

27 OpenAI et al. GPT-4 Technical Report. *Computing Research Repository*, arXiv:2303.08774, 2023. doi:10.48550/arXiv.2303.08774.

28 OpenAI et al. GPT-4o System Card. *Computing Research Repository*, arXiv:2410.21276, 2024. doi:10.48550/arXiv.2410.21276.

29 Nils Reimers and Iryna Gurevych. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, 2020. doi:10.18653/v1/2020.emnlp-main.365.

30 Cesáreo Rodríguez-Aguilera. The rise of the far right in europe. *IEMed Mediterranean Yearbook*, 2014. URL: https://www.iemed.org/publication/the-rise-of-the-far-right-in-europe/.

31 Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 399–408, 2015. doi:10.1145/2684822.2685324.

32 Javier Torregrosa, Gema Bello-Orgaz, Eugenio Martinez-Camara, Javier Del Ser, and David Camacho. A Survey on Extremism Analysis using Natural Language Processing. *Computing Research Repository*, arXiv:2104.04069, 2021. doi:10.48550/arXiv.2104.04069.

33 Mattias Wahlström and Anton Törnberg. Social Media Mechanisms for Right-Wing Political Violence in the 21st Century: Discursive Opportunities, Group Dynamics, and Co-Ordination. *Terrorism and Political Violence*, 33(4):766–787, 2021. doi:10.1080/09546553.2019.1586676.