A Transformer-Based Approach to Analyzing Public Opinion and Political Trends

Jon Gardeazabal Gutiérrez[†,]
Bilbao, Spain
jon.gargut@gmail.com

Miguel Fernandez-de-Retana[†,1,2,]
[1]*Modeling & Simulation in Life and Materials Sciences, Basque Center for Applied Mathematics (BCAM)*
[2]*Faculty of Engineering, University of Deusto*
Bilbao, Spain
mfernandez@bcamath.org

Aritz Bilbao-Jayo
Faculty of Engineering
University of Deusto
Bilbao, Spain
aritzbilbao@deusto.es

***Abstract*—Podcasts have become a significant platform for political discussion and a reflection of public opinion. This paper details the development of an NLP-driven tool designed to analyze political discourse in podcasts, with applications in smart city governance and public opinion research. The tool employs automated transcription and a RoBERTa-based classification model, trained on the Manifesto Project dataset, to categorize political topics. BERTopic is used for topic modeling, providing a structured overview of key themes. A use case illustrates the tool's effectiveness in extracting dominant political topics from podcast episodes, demonstrating its potential to provide valuable insights for urban policy and public engagement.**

***Keywords*—Political discourse, content analysis, natural language processing, public opinion**

I. Introduction

In recent years, podcasting has experienced explosive growth, becoming a global phenomenon that shows no signs of slowing down. This surge is reflected in the increasing number of listeners worldwide on all platforms that offer this service, reaching 464.7 million active listeners in 2023, a 69.1% increase from 274.8 million in 2019, and is expected to reach 504.9 million listeners in 2024 [1]. With regard to the variety of topics covered in the podcasting world, the range is extremely wide, with topics such as comedy, entertainment, and politics standing out. A survey conducted by the Pew Research Center [2] in 2022 found that 41% of the more than 5,000 adults in the United States the United States surveyed listened to podcasts about political or governmental topics.

This is precisely where the interest of this research work lies. The field of podcasting has gradually become present in terms of media. In this sense, we will carry out an analysis of the political topics discussed in these digital-era media. Therefore, we will focus on creating a tool capable of displaying the political themes based on textual data. To do so, we will employ a multidisciplinary approach to build a text categorization classifier. This classifier will be developed using a publicly available dataset of manually annotated political manifestos by the Manifesto Project [3], integrating the political science expertise of social scientists involved in the annotation process with natural language processing methodologies. This combination will allow us to automatically process substantial volumes of data.

Specifically, the initial stage will require the extraction of textual content from the audio of podcast episodes, using transcription tools such as Whisper [4]. Following transcription, we will develop and evaluate political discourse classification models to analyze the key political themes presented in these podcasts. This task will involve a comparative analysis of various text representation techniques and text classification models such as RoBERTa [5] and DistilBERT [6]. Moreover, we will employ BERTopic [7], a *state-of-the-art* topic modeling technique that combines transformer-based embeddings with class-informed term weighting to provide a structured view of the political discourse, highlighting the main thematic areas and their relationships. Finally, we will present a use case scenario of the methodology for automatic political trend analysis in podcasts.

This paper is organized as follows. Section II provides an overview of previous related work on automatic political trend analysis and the use of political manifestos as a foundation for these types of studies. Section III details our research methodology, including the dataset and the natural language processing techniques employed. Section IV explains the evaluation process of the developed classifier and presents the results. In Section V, we demonstrate a practical use case of the proposed approach by analyzing a podcast. Finally, Section VI concludes the paper and proposes directions for future research.

II. Related work

This section reviews related work in political discourse analysis, particularly content analysis techniques and their application to social media and other political texts. We highlight the transition from manual coding to automated machine learning approaches, as this area is the basis for our novel application of these techniques to podcast analysis, a relatively unexplored area to the best of our knowledge.

1) Manual Content Analysis of Political Discourse: Traditionally, political science researchers have relied on manual content analysis to study political discourse. This involves

[†]Jon Gardeazabal Gutiérrez and Miguel Fernandez-de-Retana contributed equally to this work.

human coders analyzing text (e.g., speeches, manifestos, social media posts) and assigning them to predefined categories. This approach, while providing in-depth analysis, is time consuming and expensive, especially for large datasets.

While manual content analysis of political communication was originally developed for analyzing political manifestos, researchers have adapted it for manual content analysis of political communication on newer platforms like Twitter. NLP has enabled the automation of content analysis, allowing researchers to process large volumes of text efficiently. This shift has been particularly evident in the analysis of social media data or new ways of communication such as podcasts, where manual coding is impractical. Stier et al. [8] analyzed the 2013 German federal election campaign on Twitter and Facebook, comparing topics discussed by politicians with the electorate's priorities using a Bayesian language model and german coding schema. Yaqub et al. [9] analyzed 2016 US presidential elections' discourse on Twitter, examining both public opinion and candidate sentiment.

The Manifesto Project, with its comprehensive coding scheme for political manifestos, has become a valuable resource for training NLP models to analyze a variety of political texts beyond manifestos themselves. Nanni et al. [10] used annotated political manifestos and speeches to analyze US presidential campaign speeches, applying the seven main political domains defined by the Manifesto Project. Bilbao-Jayo and Almeida [11] analyzed Spanish election discourse on Twitter using a model trained on political manifestos and a simplified political message taxonomy. They also analyzed the 2016 United States elections using all categories available in the Manifesto Project [12]. The application of the Manifesto Project's taxonomy extends beyond Twitter. For instance, Nanni et al. [13] used English political manifestos to measure Euroscepticism in transcripts of European Parliament speeches. Concurrently, significant research effort is being directed towards improving the accuracy and efficiency of manifesto classification itself. Barzallo et al. [14] developed a RoBERTa-based classifier specifically for this task. The Manifesto Project itself also conducts ongoing research [15], applying RoBERTa-based models to analyze manifestos across a multitude of languages.

These studies demonstrate the increasing use of machine learning, and particularly the use of annotated political manifestos, to automate and scale the analysis of political discourse across various platforms. Our work builds upon this trend by, leveraging recent advancements in transcription models like Whisper, introducing a novel pipeline for analyzing political discourse in podcasts and videos. This pipeline combines accurate audio transcription with a RoBERTa model fine-tuned on the Manifesto Project dataset for topic classification, enabling a comprehensive content analysis. Furthermore, our approach incorporates an interpretability analysis using BERTopic, providing insights into the resulting topic distributions.

TABLE I
CATEGORIES IN SEVEN POLICY DOMAINS [3]

Domain 1: External Relations	410 Economic Growth
101 Foreign Special Relationships: Positive	411 Technology and Infrastructure: Positive
102 Foreign Special Relationships: Negative	412 Controlled Economy: Positive
103 Anti-Imperialism: Positive	413 Nationalisation: Positive
104 Military: Positive	414 Economic Orthodoxy: Positive
105 Military: Negative	415 Marxist Analysis: Positive
106 Peace: Positive	416 Anti-Growth Economy: Positive
107 Internationalism: Positive	**Domain 5: Welfare and Quality of Life**
108 European Integration: Positive	501 Environmental Protection: Positive
109 Internationalism: Negative	502 Culture: Positive
110 European Integration: Negative	503 Equality: Positive
Domain 2: Freedom and Democracy	504 Welfare State Expansion
201 Freedom and Human Rights: Positive	505 Welfare State Limitation
202 Democracy	506 Education Expansion
203 Constitutionalism: Positive	507 Education Limitation
204 Constitutionalism: Negative	**Domain 6: Fabric of Society**
Domain 3: Political System	601 National Way of Life: Positive
301 Decentralisation: Positive	602 National Way of Life: Negative
302 Centralisation: Positive	603 Traditional Morality: Positive
303 Govern. and Admin. Efficiency	604 Traditional Morality: Negative
304 Political Corruption: Negative	605 Law and Order
305 Political Authority: Positive	606 Civic Mindedness: Positive
Domain 4: Economy	607 Multiculturalism: Positive
401 Free-Market Economy: Positive	608 Multiculturalism: Negative
402 Incentives: Positive	**Domain 7: Social Groups**
403 Market Regulation: Positive	701 Labour Groups: Positive
404 Economic Planning: Positive	702 Labour Groups: Negative
405 Corporatism: Positive	703 Agriculture and Farmers
406 Protectionism: Positive	704 Middle Class and Professional Groups: Positive
407 Protectionism: Negative	705 Minority Groups: Positive
408 Economic Goals	706 Non-Economic Demographic Groups: Positive
409 Keynesian Demand Management: Positive	

III. RESEARCH METHODOLOGY

This section details the technical development of this research, focusing on the dataset, used text representation and classification models, as well as the definition of the tasks we aim to solve. The code is available on GitHub [1].

A. Electoral Program Classification

The dataset of political manifestos used in this research is the public Comparative Manifestos Project (CMP) dataset. Today, the categorization schema used by the CMP for the annotations of political manifestos consists of 56 main categories (see Table I). The annotation of manifestos involves two key steps: segmentation into coding units and category assignment. Since a single sentence may express multiple ideas, the text is first divided into "quasi-sentences," each representing a distinct statement. Subsequently, each of these quasi-sentences is assigned a category. The dastaset used in this manuscript consists of 115,305 quasi-sentences from English-language electoral programs, labeled according to previously mentioned taxonomy which has been widely used by political scientists.

For model training and evaluation, the dataset is split into training, evaluation, and test sets. To ensure consistent evaluation, the same data splits are used in all models.

In this work, we explore several NLP techniques for text representation and classification. For text representation, we employed Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and BERT [16] embeddings. TF-IDF, a weighted word representation, was used to quantify word importance based on frequency and specificity within the corpus. Word2Vec, a noncontextual method, provided semantic knowledge by training a neural network to predict neighboring words. BERT embeddings, a contextual representation method,

[1] https://github.com/Jongarde/TFM

captured a broader semantic understanding by considering the surrounding context. These different representation methods allow us to compare the impact of contextual and non-contextual embeddings in our classification task.

For classification, we utilized both machine learning and deep learning models. The machine learning models included Logistic Regression, Histogram-Based Gradient Boosting (HBGB), and Support Vector Machines (SVM). These models were trained from scratch on our manifestos dataset. The deep learning models comprised Convolutional Neural Networks (CNN) and fine-tuned BERT architecture models (RoBERTa and DistilBERT). The CNN architecture leverages n-grams to capture text sequence features. The BERT-based models were fine-tuned for our specific classification task, using the pre-trained models as feature extractors. We also employ BERTopic, a topic modeling technique, in a supervised manner, using our labeled data to guide topic extraction and enhance the interpretability of our classification results.

B. Supervised Topic Modeling

In addition to the classification models used to identify and predict political topics, we also explored the use of a (supervised) *topic modeling* approach to extract the main topics from our corpus. In essence, as opposed to traditional (*unsupervised*) topic modeling methods where latent human-interpretable thematic structures are uncovered probabilistically without leveraging any prior knowledge [17]; supervised topic modeling incorporates label information to guide topic discovery, aligning results with predefined categories. In this work, we employ BERTopic [7], a neural framework that combines transformer-based embeddings with class-informed term weighting, to analyze political discourse within a supervised scenario. At its core, BERTopic operates through four stages:

1) *Embedding:* Documents are first encoded into dense vector representations using Sentence-BERT (SBERT) [18]. SBERT is employed to extract semantically rich embeddings through *contrastive learning*. In essence, the model is trained to learn sentence embeddings by contrasting similar examples against dissimilar ones, effectively pulling semantically similar inputs closer in vector space while pushing dissimilar ones apart. However, this latent semantically-rich representation is used *only* to cluster semantically similar documents and not directly in generating the actual topics.
2) *Dimensionality Reduction:* UMAP [19] projects embeddings into a lower-dimensional space while preserving local and global structures.
3) *Clustering:* Then, the *low-dimensional* embeddings are aggregated into clusters (i.e., our latent topics) using the hierarchical density-based HDBSCAN method.
4) *Topic Representation:* Finally, class-based TF-IDF (cTF-IDF) extracts discriminative terms for each cluster (i.e., topic), weighted by their frequency within a class relative to others.

In the supervised variant, label information directly informs the cTF-IDF computation in (1). Instead of deriving clusters purely from embeddings, predefined classes (i.e., the 56 political topics in our dataset) serve as *pseudo*-clusters. In practice, steps *2)* and *3)* above are replaced by a classification model. cTF-IDF then calculates term importance for each class c as:

$$\text{cTF-IDF}(t, c) = \| \operatorname{tf}(t, c) \| \cdot \log \left(1 + \frac{N}{\operatorname{df}(t, C)} \right) \quad (1)$$

where $\operatorname{tf}(t, c)$ is term t's frequency in class c, N is the average number of words per class, and $\operatorname{df}(t, C)$ is the number of classes containing t. Note that the term-frequency is L_1-normalized to account for differences in topic sizes.

The use of cTF-IDF in this supervised context accentuates the most salient terms (*keywords*) within each labeled group enhancing the interpretability of the topics: we can extract the words that give a good representation of the input classes and use it as validation of the features used in the classification models. Both variants of BERTopic are summarized in Fig. 1.

IV. Evaluation

We evaluated the classification models using several key metrics. Accuracy measures the overall proportion of correct predictions. Top-k Accuracy assesses whether the true class is within the model's top k predicted probabilities, important for our multi-class problem. Macro-averaged Precision, Recall, and F_1-score were computed, averaging the per-class metrics to give equal weight to all 56 topics. Moreover, we also used the Matthews correlation coefficient (MCC). Unlike F_1-score, MCC is invariant to class swapping and considers all four values of the confusion matrix, making it robust to class imbalance, a known issue with the Manifesto Project dataset. MCC ranges from -1 (perfectly inverse) to +1 (perfect), with 0 representing an average random prediction.

Regarding machine learning models, two different types of results are obtained as it can be seen in Table II: those offered by TF-IDF and those using BERT embeddings. The latter results are significantly inferior to the former. This decrease in performance may be due to these models' misunderstanding of the semantic knowledge provided by BERT embeddings.

However, the poor performance of machine learning models with BERT embeddings is not generalized to other approaches with the same type of models. In fact, models such as logistic regression or SVM show an improvement in results compared to deep learning approaches, such as CNN. In fact, the SVM method comes significantly close in performance to transformer-based models. A surprising result is the precision obtained by the SVM method, which leads the ranking by a wide margin, even including the most powerful models used in this project, such as those based on the BERT architecture.

Finally, the two RoBERTa and DistilBERT models coincide with the two best results obtained. Likewise, the RoBERTa model offers better results than DistilBERT. This was expected for two reasons: first, it is necessary to remember that DistilBERT is nothing more than the traditional BERT model subjected to a knowledge distillation process. This consequently reduces the NLU capabilities of the model and, therefore, slightly worsens the model's results. Secondly, it

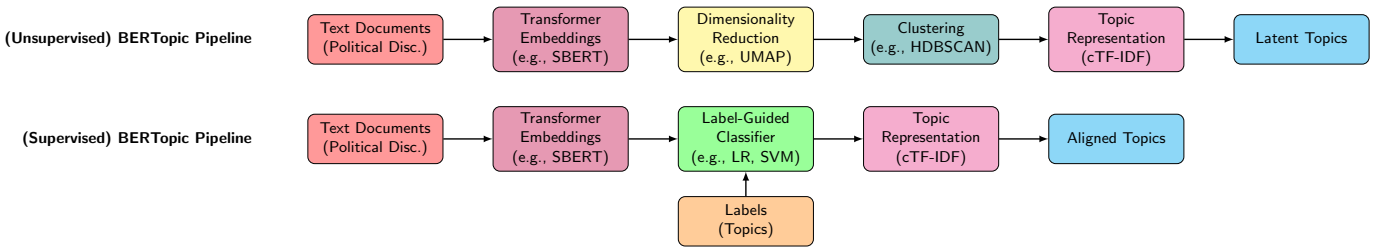

Fig. 1. Pipelines for BERTopic (Unsupervised & Supervised Versions)

TABLE II
CLASSIFICATION RESULTS

Model	Accuracy	Acc.@2	Acc.@3	Acc.@5	Recall	F_1 Score	Precision	MCC
Logistic Regression (TF-IDF)	0.4977	0.6492	0.7311	0.8172	0.3324	0.3551	0.4204	0.4716
Hist Gradient Boosting (TF-IDF)	0.2814	0.4137	0.5007	0.6079	0.1731	0.1586	0.1857	0.2446
Support Vector Machine (TF-IDF)	0.4392	0.6135	0.6971	0.7899	0.1945	0.1907	0.2407	0.4074
CNN (BERT Embeddings)	0.5516	0.6870	0.7546	0.8289	0.3448	0.3590	0.3975	0.5286
RoBERTa	0.6509	0.7824	0.8386	0.8980	0.4594	0.4699	0.5204	0.6334
DistilBERT	0.6180	0.7575	0.8223	0.8851	0.4049	0.4099	0.4466	0.5990

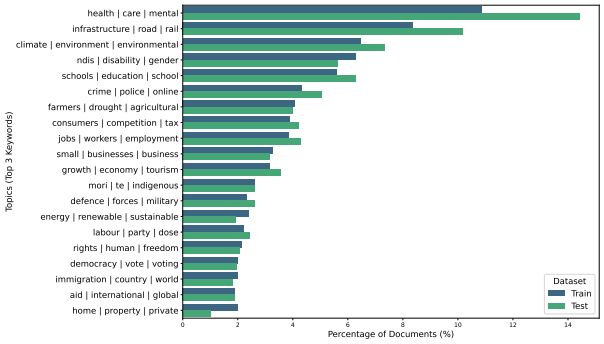

Fig. 2. Topic Distribution (Relative Frequencies for Top 20 Topics)

must also be recognized that RoBERTa's pre-training is more complex than BERT's, which, unlike the previous model, allows for improved NLU capabilities. Due to these results in the model's precision, the RoBERTa model is chosen to face the classification problem in the use case.

Regarding the supervised topic modeling, the results using BERTopic in combination with a Support Vector Machine (SVM) classifier (the *shallow* classifier with the best classification results in Table II) to analyze the political topics in the dataset are presented in Figures 2, 3, and 4. The results show a clear distribution of topics, with some topics being more prevalent than others. The topic hierarchy provides a structured view of the political discourse, revealing the relationships between the topics. Finally, the top topics with keyword scores highlight the most important topics in the dataset and provide rich insight into the actual political message.

First, the topic distribution comparison in Fig. 2 shows the relative frequencies of the (top-20) topics identified by BERTopic in the dataset, both in the train and test cohorts. As expected, the distribution is similar in both sets, where pronounced *inter-topical* variations are observed: some topics,

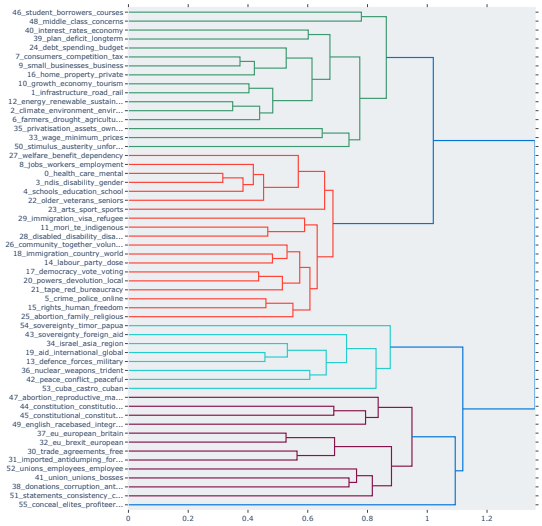

Fig. 3. Political Topic Hierarchy Using Ward's Hierarchical Clustering

specifically those related to economic and social issues, are more prevalent than others. In fact, these overrepresented topics reflect the current focus of political discourse on themes such as mental health, education, gender equality, public infrastructure, security, employment, or the environment.

In contrast, the hierarchy of topics in Fig. 3 illustrates the relationships between these topics, revealing a more structured view of the political discourse. The hierarchy is constructed using Ward's method [20] based on the euclidean similarity between the topics. Observing the main branches, we can identify four potential groupings suggesting a cluster of closely-related topics. Top to bottom, these broadly relate to:

1) *Economics & Public Finances:* First, the top branch encompasses topics concerning economic and financial issues, such as public spending, taxation, and economic

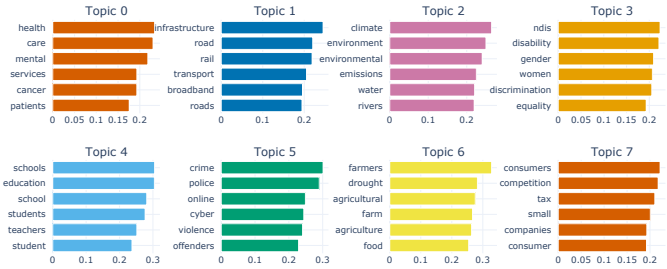

Fig. 4. cTF-IDF Keyword Scores for the Top 8 Most Prevalent Topics

growth. This group also includes topics related to public infrastructure, transportation, private property, and even sustainable development.

2) *Social & Community Welfare:* The second branch focuses on social and community welfare topics, including education, employment, (mental) health, social well-being, immigration, or even arts and culture. These are all particularly relevant in the context of social policies and community development.
3) *National Sovereignty & International Relations:* Next, our third branch is centered around matters related to national sovereignty and international relations. This includes topics such as security, defense, foreign policy, and international cooperation. Likewise, international conflicts and peacekeeping are a key part of this group. These topics are crucial in shaping a country's foreign policy and its role in the international community.
4) *Political & Economic Governance:* Finally, this is the most *heterogeneous* group, encompassing topics related to economic and political governance with a focus on transparency, national interest, labor, and constitutional/individual rights. It is a mix of economic policy (trade, unions), national sovereignty (EU/Brexit), ethical governance (corruption, transparency), and individual rights (abortion, integration and constitutionalism).

Of course, these groupings are far from exhaustive. Nonetheless, they provide a structured view of the political discourse, highlighting the main thematic areas and their relationships. In practice, this can be particularly useful for understanding the main political trends, as well as for identifying potential areas of interest in the discourse of political actors. Interestingly, the topics identified by BERTopic are very consistent. For instance, one may observe that topics related to abortion are included in both the second and fourth branches. In fact, if we look closely, the topic included in the second branch is more related to the social implications of abortion (from a more traditional perspective), while the topic in the fourth branch is more related to the political and ethical implications of abortion and the regulatory framework surrounding it. By using BERTopic we are able to analyze and understand the context in which topics are discussed. Moreover, given these *high-level* groupings, we can compare them to the *super-*domains in the Manifesto Project dataset as in Table III.

Finally, Fig. 4 illustrates the top topics with their keyword scores. For each significant topic, the visualization lists characteristic keywords, revealing semantic themes. For instance, Topic 0 with keywords like "health," "care," and "cancer" indicate a focus on health services, while Topic 1 with "infrastructure," "road," and "rail" represents infrastructure projects, and so on for other topics covering environment, social inclusion, education, crime and law enforcement, agriculture, and consumer rights and economic regulation, respectively. These keywords provide interpretable insights into the concrete and key concepts driving each political topic, offering comprehensive and multi-faceted analysis of the discourse, and the relationships between the various topics.

In conclusion, the analysis of political discourse using BERTopic unveils a landscape dominated by specific thematic areas. The most present topics within the corpus are demonstrably those related to economic and social issues. As evidenced by the topic distribution and keyword analysis, concerns around themes such as mental health, education, gender equality, public infrastructure, security, employment, and the environment are particularly salient. These topics, clustered into the domains of Economics & Public Finances and Social & Community Welfare, suggest a significant emphasis in contemporary politics on socio-economic well-being, community development, and foundational public services.

V. Use Case Scenario: Podcast Analysis

This section demonstrates the practical application of our system to analyze the political content of a podcast episode. The analysis pipeline began with extracting the audio track from the podcast. The audio was then transcribed into text with the faster-whisper model. The resulting transcript was segmented into individual sentences, followed by a preprocessing step to remove short sentences, ensuring only relevant political discourse remained. Finally, cleaned sentences were classified into the 56 topic categories using our fine-tuned RoBERTa model, and the frequency of each predicted topic was calculated to determine the dominant themes.

The episode analyzed exhibited a strong prevalence of the topic "Political Authority (305)", accounting for more than 50% of the classified sentences. "Democracy (202)" was the second most frequent topic. This distribution differed significantly from the class distribution in the training data, where these two topics are not among the most frequent. This highlights the model's ability to identify dominant themes even when they are not overrepresented in the training set, demonstrating robustness to class imbalance. The top five topics are: Political Authority, Democracy, Law and Order, Equality: Positive, National Way of Life: Positive.

VI. Conclusions and Future Work

The study effectively demonstrates a pipeline for analyzing political discourse in podcasts, combining accurate audio transcription with a RoBERTa model fine-tuned on the Manifesto Project dataset. Additionally, the study utilizes BERTopic for topic modeling, providing a structured view of the political discourse and highlighting the main thematic areas. For future

TABLE III
RELATIONSHIPS BETWEEN BERTOPIC CLUSTERS & MANIFESTO PROJECT DOMAINS

BERTopic Cluster	Strongly Represented	Moderately Represented	Weakly Represented	Not Represented
Economics & Public Finances	D.4: Economy	D.3: Political System, D.5: Welfare & QoL	D.2: Freedom & Dem., D.7: Social Groups	D.1: Ext. Relations, D.6: Social Fabric
Social & Community Welfare	D.5: Welfare & QoL	D.2: Freedom & Dem., D.4: Economy, D.6: Social Fabric, D.7: Social Groups	D.1: Ext. Relations, D.3: Political System	None
National Sovereignty & Intl. Rel.	D.1: Ext. Relations	D.3: Political System, D.4: Economy	D.2: Freedom & Dem., D.5: Welfare & QoL, D.6: Social Fabric	D.7: Social Groups
Political & Economic Governance	D.2: Freedom & Dem., D.3: Political System, D.4: Economy	D.1: Ext. Relations, D.6: Social Fabric, D.7: Social Groups	D.5: Welfare & QoL	None

work, Large Language Models (LLMs) with few-shot approaches could be explored. Moreover, to improve the analysis of podcast conversations, future research could also focus on improving diarization techniques, enabling the transcription to be divided by speaker. This would allow for the visualization of who said what in the podcast episodes,allowing new studies that analyze the conversation by each speaker.

ACKNOWLEDGMENT

We gratefully acknowledge the support of the Ministry of Economy, Industry, and Competitiveness of Spain under Grant No.:INCEPTION (PID2021- 128969OB-I00) and the Basque Government under the grant DEUSTEK5 (IT1582-22).

REFERENCES

[1] "13 Podcast Statistics You Need To Know For 2024 — backlinko.com." https://backlinko.com/podcast-stats. [Accessed 28-02-2025].

[2] E. Shearer, J. Liedke, K. E. Matsa, M. Lipka, and M. Jurkowitz, "Podcasts as a Source of News and Information." https://www.pewresearch.org/journalism/2023/04/18/podcasts-as-a-source-of-news-and-information/, 2023. [Accessed 28-02-2025].

[3] A. Volkens, W. Krause, P. Lehmann, T. MatthieÃŸ, N. Merz, S. Regel, and B. WeÃŸels, "The manifesto data collection. manifesto project (mrg/cmp/marpor). version 2018b," 2018.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, pp. 28492–28518, PMLR, 2023.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[7] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[8] S. Stier, A. Bleier, H. Lietz, and M. Strohmaier, "Election campaigning on social media: Politicians, audiences, and the mediation of political communication on facebook and twitter," *Political communication*, vol. 35, no. 1, pp. 50–74, 2018.

[9] U. Yaqub, S. A. Chun, V. Atluri, and J. Vaidya, "Analysis of political discourse on twitter in the context of the 2016 us presidential elections," *Government Information Quarterly*, vol. 34, no. 4, pp. 613–626, 2017.

[10] F. Nanni, C. Zirn, G. Glavas, J. Eichorst, and S. P. Ponzetto, "Topfish: Topic-based analysis of political position in us electoral campaigns," 2016.

[11] A. Bilbao Jayo and A. Almeida, "Political discourse classification in social networks using context sensitive convolutional neural networks," in *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, (Melbourne, Australia), pp. 76–85, July 2018.

[12] A. Bilbao-Jayo and A. Almeida, "Improving political discourse analysis on twitter with context analysis," *IEEE Access*, vol. 9, pp. 104846–104863, 2021.

[13] F. Nanni, G. Glavaš, S. P. Ponzetto, S. Tonelli, N. Conti, A. Aker, A. P. Aprosio, A. Bleier, B. Carlotti, T. Gessler, *et al.*, "Findings from the hackathon on understanding euroscepticism through the lens of textual data," LREC, 2018.

[14] F. Barzallo, M. E. Moscoso, M. Pérez, M. Baldeon-Calisto, D. Navarrete, D. Riofrío, P. Medina-Pérez, and S. K. Lai-Yuen, "A zoom into ecuadorian politics: Manifesto text classification using nlp," in *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*, pp. 1–6, IEEE, 2023.

[15] T. Burst, P. Lehmann, S. Franzmann, D. Al-Gaddooa, C. Ivanusch, S. Regel, F. RiethmÃ¼ller, B. WeÃŸels, and L. Zehnter, "manifesto-berta. version 56topics.sentence.2024.1.1," 2024.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

[17] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Information Systems*, vol. 112, p. 102131, 2023.

[18] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[19] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[20] F. Murtagh, "Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm," *ArXiv abs/1111.6285*, 2011.