Enhancing Political Topic Detection: Chaining Automated Speech Recognition and Language Models for Automated Classification

Mariana S. Ramos[1[0000-0001-9535-6445]], Pedro Chaves[2[0000-0003-4201-5641]], Pedro R. Almeida[1], Patrício Costa[1,3,4[0000-0002-1201-9177]], Pedro Silva Moreira[1,5[0000-0002-2800-3903]]

[1] MindProber Labs
[2] Select Data
[3] Faculty of Psychology and Education Sciences, University of Porto, Porto, Portugal
[4] Center for Psychology at University of Porto, Porto, Portugal
[5] Psychology Research Centre, School of Psychology, University of Minho, Braga, Portugal

lncs@springer.com

Abstract In this work, we present an approach for the automatic classification of political topics in the context of TV-broadcasted political debates. The full recordings of five Republican Party political debates were subjected to a pipeline involving automatic speech recognition and speaker diarization. The output chunks were then automatically classified based on a set of predefined political topics, according to (1) natural language processing (NLP, using mDeBERTa) and (2) large language models (LLMs, using GPT-4o, llama3-8b and llama 3-70b). The performance of the models was compared against manual classification. The results demonstrated that GPT-4o had the highest accuracy (69%) followed by llama3-70b (67%), llama3-8b (61%) and mDeBERTa (43%). Models' accuracy further improved when considering secondary manual classifications from the human coders (GPT-4o: 75%, llama3-70: 74%, llama3-8b: 69%, mDeBERTa: 43%). This research demonstrates the viability of automated text classification, based on LLMs to summarize political debates.

Keywords: political debates, automatic speech recognition, speakers diarization, natural language processing, large language models

1 Introduction

Automated text analysis has emerged as a promising tool for complementing traditional survey-based methods, particularly within the realm of political science [1]. Several approaches have been studied, notably automatic identification of political viewpoints and hate speech [2, 3].

Analyzing political debates between candidates is a crucial tool for understanding the political landscape. This paper suggests automating the extraction of speeches, speakers, and the topics discussed through a pipeline chaining automated speech recognition, diarization and topic extraction. Through the integration of speech, speaker, and topic, we glean valuable insights into the perspectives of each candidate and, potentially, their prominent positions in specific domains. For this purpose, we focus on a sequence of recent debates of the Republican party for the nomination of the candidate for president of the United States.

This approach allows for a more granular and detailed analysis of political debates. Furthermore, by creating a dictionary of topics, it becomes feasible to categorize and organize the discourse more effectively, facilitating the identification of patterns, trends, and gaps in the discussions. Thus, this methodology not only helps elucidate the differences among candidates but also can be used to assess the representativeness and scope of the issues discussed during the electoral campaign.

1.1 **Related Work**

The field of speech recognition has attracted considerable attention in the last decade, driven by the availability of powerful computing resources and large amounts of training data using different approaches and technologies [Fraihat2024].

Traditional speech recognition methods encompass Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs). HMMs utilize statistical Markov processes to model the probabilistic sequence of spoken words, crucial for temporal pattern recognition in audio [rabiner1989]. Conversely, GMMs are employed to model probability distributions of observed data, often used in acoustic modeling to represent features such as frequency spectra [reynolds1995]. While foundational in early ASR (Automatic Speech Recognition) systems, these methods are increasingly supplanted by deep learning approaches, which excel in capturing complexities and enhancing overall recognition accuracy.

Deep learning-based methods have revolutionized the field of speech recognition by significantly enhancing accuracy and versatility. These methods employ sophisticated neural network architectures tailored to handle the complexities of speech data. Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks, are pivotal for their ability to model temporal dependencies within speech sequences, enabling them to maintain context over extended periods. Convolutional Neural Networks (CNNs), originally designed for image processing, have been adapted to extract intricate acoustic features from audio signals, making them invaluable in speech recognition tasks. Moreover, Transformer models have emerged as powerful alternatives, utilizing self-attention mechanisms to capture nuanced relationships across input sequences. Models like Whisper and wav2vec 2.0 showcase the effectiveness of Transformer architectures in achieving state-of-the-art results by efficiently processing and understanding speech data. These advancements underscore the ongoing evolution towards more accurate and robust automatic speech

recognition systems, driven by deep learning innovations [Prabhavalkar2024, Latif12023].

While traditional models like HMMs and GMMs laid the groundwork for ASR, deep learning methods have demonstrated superior performance in capturing the complexities of speech data. Innovations such as Whisper highlight the potential of self-supervised methods in reducing the reliance on extensive labeled data, further promoting advancements in the field.

Similar to speech recognition, the field of speaker diarization has also gained significant traction in recent years, driven by advances in machine learning techniques and the increasing availability of large annotated datasets. Speaker diarization involves the process of segmenting and grouping an audio recording based on different speakers. Traditional approaches to speaker diarization often relied on statistical models and heuristics to distinguish between speakers based on acoustic characteristics and temporal patterns.

Speaker diarization has traditionally relied on clustering-based methods, Gaussian Mixture Models (GMMs), and Hidden Markov Models (HMMs), which focus on segmenting and grouping audio based on acoustic features and probabilistic modeling. Clustering algorithms and GMMs model the distribution of speech features, while HMMs capture the sequential nature of speech to identify speaker transitions. In contrast, modern deep learning methods have significantly advanced the field by leveraging sophisticated neural network architectures. Recurrent Neural Networks (RNNs) and their variants, like Long Short-Term Memory (LSTM) networks, excel in modeling temporal dependencies in speech data, whereas Convolutional Neural Networks (CNNs) extract high-level acoustic features. Transformer models, with their self-attention mechanisms, handle complex speaker dynamics and overlapping speech more effectively, while End-to-End systems streamline the diarization process by directly predicting speaker boundaries and labels from raw audio data. These advancements represent a shift towards more accurate and efficient speaker diarization technologies**.**

2 Methods

2.1 Data

Five Republican Party political debates conducted in the United States of America were utilized in this study. We selected this set of debates as a proof of concept for this pipeline. Table 1 details the debates to be analyzed, including the date, channel and the names of moderators and candidates.

Table 1. Date, channel, moderators and candidates of debates selected for analysis.

Debate	Date	Channel	Moderators	Candidates
First GOP Debate	2023/08/23	FOX News	Bret Baier Martha	Asa Hutchinson Chris Christie

			MacCallum	Vivek Ramaswamy Doug Burgum Mike Pence Nikki Haley Ron DeSantis Tim Scott
Second GOP Debate	2023/09/27	FOX News	Stuart Varney Dana Perino Ilia Calderón	Chris Christie Doug Burgum Mike Pence Nikki Haley Tim Scott Vivek Ramaswamy Ron DeSantis
Third GOP Debate	2023/11/08	NBC	Lester Holt Kristen Welker Hugh Hewitt	Chris Christie Nikki Haley Ron DeSantis Vivek Ramaswamy Tim Scott
Fourth GOP Debate	2023/12/06	NewsNation	Megyn Kelly Elizabeth Vargas Eliana Johnson	Chris Christie Ron DeSantis Vivek Ramaswamy Nikki Haley
Republican Presidential Primary Debate	2024/01/10	CNN	Jake Tapper Dana Bash	Nikki Haley Ron DeSantis

All debates lasted approximately 2 hours, with periods of commercial activity that varied between 11 and 17 minutes. Next, Table 2 specifies the duration of each debate, including the period of commercial activity also detailed in the same table.

Table 2. Total duration and duration of commercial activity of each debate.

Debate	Total duration	Duration of commercial activity
First GOP Debate	02h02min	17min
Second GOP Debate	02h00min	16min
Third GOP Debate	01h57min	11min
Fourth GOP Debate	01h59min	17min
Republican Presidential Primary Debate	02h06min	13min

2.2 Automatic speech recognition

Automatic speech recognition (ASR) is capable of converting spoken language into written text. In this study, Whisper was the chosen method due to its robustness and ability to handle a variety of challenges in automatic speech recognition. Its effectiveness in accurately transcribing different languages, dealing with diverse accents, background noise, and specialized technical language made it an ideal choice for the project [4].

Furthermore, Whisper has been trained on a massive dataset consisting of 680 000 hours of multilingual and diverse supervised data sourced from the web. This extensive training corpus has enabled Whisper to learn from a wide range of linguistic variations and contexts, contributing to its robustness and effectiveness in various speech recognition tasks [4].

The Whisper architecture, represented in Fig.1, incorporates a seamless end-to-end approach, adopting an encoder-decoder transformer paradigm. Initially, the input audio is segmented into 30-second chunks, which are transformed into log-Mel spectrograms. These spectrograms are then processed by an encoder module. A decoder is trained to predict the corresponding text caption. Embedded in this process are specialized tokens that model a wide range of tasks, namely, language identification, sentence-level timestamps, multilingual speech transcription, and speech-to-English translation [4].

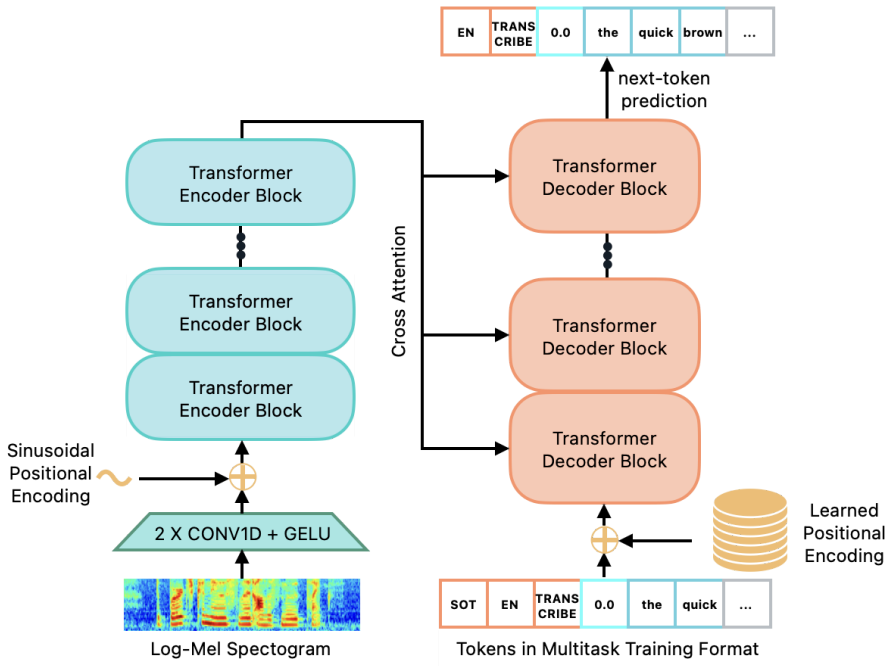

Fig. 1. The Whisper architecture implemented as an encoder-decoder transformer. Adapted from [4].

ASR was implemented using an implementation of Whisper Large V3, Insanely Fast Whisper (https://github.com/Vaibhavs10/insanely-fast-whisper), which was developed to achieve fast performance of the transcriptions. The pipeline enables the extraction of timestamps for words and chunks (defined as periods between pauses). Here, we opted for the maximization of the temporal granularity of the transcriptions, by extracting the timestamps for each word. This approach was relevant to match the transcription with the diarization, in which different speakers are frequently identified in the same chunk. Due to the computational demands of this task, this implementation was performed with Replicate (https://replicate.com/), an approach for running machine learning tasks on the cloud.

2.3 Speaker diarization

Speaker diarization is the task of partitioning an audio stream into homogeneous temporal segments according to the speaker's identity [5]. To obtain speaker diarization, an open source toolkit *pyannote.audio* (version 3.1) was used. *pyannote.audio* consist in the following steps [6]:

(1) speaker segmentation applied to a short sliding window
(2) neural speaker embedding of each (local) speakers
(3) (global) agglomerative clustering

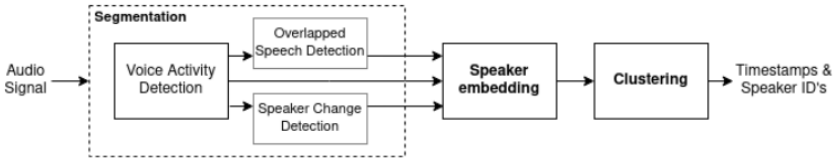

Fig. 2. Diarization system based on PyAnnote [7].

2.4 Speech and speakers aggregation

The integration of speech detected through Whisper with speaker diarization is highly significant, as it allows for extracting meaningful information from conversational content.

When employing speech recognition and diarization as separate methods, inconsistencies in timing occasionally arose during the data aggregation process. To avoid this type of error, the data was analyzed to the hundredth of a second for each word and chunk obtained through ASR, and for speakers obtained through diarization. This way, it is possible to achieve greater granularity at the word level and greater confidence in the assigned speaker. After identifying speakers, a custom preprocessing pipeline was implemented to filter the number of identified speakers. In live broadcasting, it is frequent for the voice of the intervenient to be overimposed with the sound of the audience which may lead to a mis-identification of the speaker through automated pipelines. In addition, as mentioned previously, during the debate

there is advertising content. To deal with this and to avoid the need for manual inspection of the content, the total number of words associated with each speaker was counted and only speakers with more than 200 words (about 1% of words media per debate) were considered for the analysis. The implication of this strategy with regards to data loss is summarized in Table 3.

Table 3. Initial and final number of words and speakers (before and after filtering) and data loss, in percentage.

Debate	Initial number of words	Final number of words	Initial number of speakers	Final number of speakers	Retained rate (%)
First GOP Debate	17965	15751	59	10	87.68
Second GOP Debate	17638	15581	48	10	88.34
Third GOP Debate	19806	18579	31	8	93.80
Fourth GOP Debate	20387	18092	52	7	88.74
Republican Presidential Primary Debate	21688	20592	34	4	94.95

This filtering pipeline results in only the relevant participants in the debate, the moderators and the candidates, emerging. Fig. 3 shows, as an example, a representation of the speakers at the First GOP Debate. The moderators correspond to the last two lines of the graph, and the candidates correspond to the remaining eight lines.

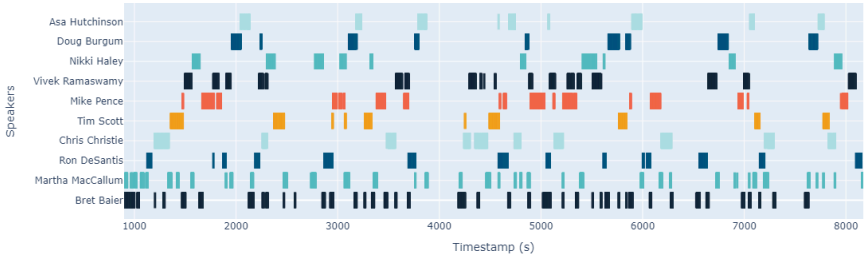

Fig. 3. Representation of speakers activity during the First GOP debate.

2.5 Topics extraction

The assignment of topics added to the speech and speakers is extremely relevant in the political context. Systems based on Artificial Intelligence (AI) are widely used today to make decisions that impact individuals and society [8]. Knowing the potential of artificial intelligence to interpret and classify data, different approaches to classifying chunks based on a topic dictionary were tested.

By definition, topic extraction is a dynamic task that can change with the scope of the content analyzed - in the specific example of political debates, it is expected that debates occurring at different periods or in different regions will tackle different aspects. To tackle this problem and the general absence of pre-trained models aimed at this task, we've conceptualized topic extraction as a zero-shot classification problem, comparing the performance of different types of language models.

2.5.1 Dictionary construction

A dictionary was created with topics tailored to politics in the United States of America. The dictionary was created by experts in the field. The dictionary includes the following topics: border control/immigration, civil rights, crime, economy, education, elections, environment, foreign policy/war, healthcare, housing, infrastructure, justice, labor, security, taxes. A codebook was created with each of the labels and their respective definitions and examples, to facilitate the assignment of topics.

2.5.2 DeBERTa

DeBERTa (Decoding-enhanced BERT with disentangled attention) is an advanced variant of BERT (Bidirectional Encoder Representations from Transformers) architecture, developed by Microsoft. The main innovation of DeBERTa is the introduction of enhanced attention and representation mechanisms that improve the model's performance in various natural language processing (NLP) tasks [9]. In this study, the improved version of this model, mDeBERTaV3, was used. This version enhances DeBERTa by employing training loss with replaced token detection (RTD) and introducing a new weight division method [10]. For this study we've leveraged a fine-tuned version of mDeBERTaV3 for zero-shot classification, which used the topic dictionary as the source of classes for the text classification [10].

2.5.3 LLM-based approaches

For LLM-based approaches, we've refined a simple standardized system prompt for the different models. The prompt was defined as: "You are an expert topic classifier which will accept a block of text, and decide where the block of text fits into. You

excel in political debate analysis. You only return a prediction and nothing else". Afterwards, for each individual block of text, the following user prompt was passed "Consider this block of text: {text}\nConsider this list of topics:{topics_list}\nReturn a single prediction where you select the best possible topic for the block of text. Return the topic name only.". This approach follows a common system-user prompt pattern and effectively implements zero-shot classification for any range of predefined topics.

2.5.4 **Llama-3**

Llama models are a family of large language models developed by Meta. Since the release of the first model from the family, in 2023, Llama has gained an important status among open-source LLMs. The recent release of Llama 3 set a new state-of-the-art (SoTA) for open-source models. Despite the fact that the technical paper has not yet been released, Llama-3 has gained immediate popularity. As other LLMs, Llama-3 uses a transformers architecture, which facilitates efficient parallelization during training and inference. Llama-3 was trained on over 15 trillion tokens, a 7-times increase compared to its predecessor Llama-2 and has a knowledge cutoff until March, 2023. The standard benchmarking reveal that Llama-3 outperforms other popular open-source LLMs, including Mistral and Gemma, in tasks such as Massive Multitask Language Understanding (MMLU) and, importantly, for the scope of this paper, on AGIEval, which includes question-answering, summarization, and sentiment analysis (https://ai.meta.com/blog/meta-llama-3). Llama inference was conducted using Groq (https://groq.com).

2.5.5 **GPT-4o**

Generative Pre-trained Transformer (GPT) is a large-scale, multimodal language model developed by OpenAI. In this study, the latest version of this model, GPT-4o (released on May 13th), was used. Despite the lack of a technical report of GPT-4o at the date of this submission, the model is based on GPT-4, a Transformer-based model aimed to predict the next token in a document [11]. It has been demonstrated that GPT-4 often outscores the vast majority of human test takers, achieving notable scores in simulated bar exams. It also has a notable performance on MMLU, obtaining a score of 86.4% [11].

2.5.6 **Manual classification**

Three different experts were chosen to manually classify the debates to serve as ground truth. For the manual classification, they were only given text chunks and the list of topics mentioned above. The process was blind to the experts, as they did not have any information about the classification made by previous algorithms or other experts. In the case of ambiguity or presence of multiple topics, the experts were asked to select a primary topic and to identify the remaining as additional topics.

3 Results

The confusion matrices for each individual model are represented on Fig. 4. Table 4 shows performance metrics for individual classes per model. The classification on war/foreign policy topics had the greatest focus during the debates and were those with the highest precision scores (91% for GPT-4o, 75% for llama3-8b, 87% for llama3-70b and 94% for mDeBERTa). Other topics with high correct hits included the topics elections and immigration/border control.

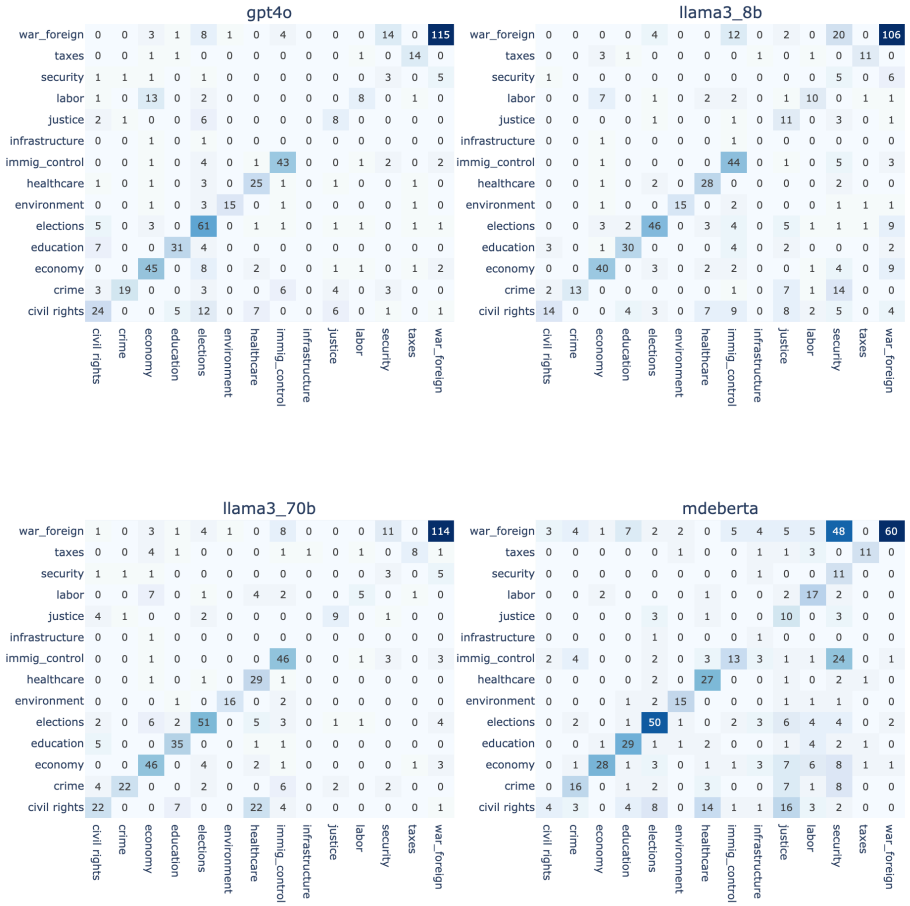

Fig. 4. Confusion matrices for each individual model.

https://doi.org/10.33774/apsa-2024-24qbn **ORCID:** https://orcid.org/0000-0002-1201-9177

Table 4. Performance metrics for individual classes by model.

	GPT-4o			llama3-8b			llama3-70b			mDeBERTa		
Topic	**prec**	**recall**	**f1-score**	**prec**	**recall**	**f1-score**	**prec**	**recall**	**f1-score**	**prec**	**recall**	**f1-score**
civil rights	0.55	**0.43**	**0.48**	**0.70**	0.25	0.37	0.56	0.39	0.46	0.44	0.07	0.12
crime	0.90	0.50	**0.64**	**1.00**	0.34	0.51	0.92	**0.58**	0.71	0.53	0.42	0.47
economy	0.64	0.73	0.68	0.69	0.65	0.67	0.66	**0.74**	**0.70**	**0.88**	0.45	0.60
education	**0.82**	0.74	0.78	0.81	0.71	0.76	0.74	**0.83**	**0.79**	0.66	0.69	0.67
elections	0.53	**0.81**	0.64	0.77	0.61	0.68	**0.78**	0.68	**0.73**	0.66	0.67	0.66
environment	0.94	0.71	0.81	**1.00**	0.71	0.83	0.94	**0.76**	**0.84**	0.75	0.71	0.73
healthcare	**0.69**	0.76	0.72	0.67	0.85	**0.75**	0.46	**0.88**	0.60	0.52	0.82	0.64
housing	0.00	0.00	0.00				0.00	0.00	0.00	0.00	0.00	0.00
immig_control	**0.77**	0.80	**0.78**	0.54	0.81	0.65	0.61	**0.85**	0.71	0.59	0.24	0.34
infrastructure	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.50	0.11
justice	0.38	0.47	0.42	0.30	**0.65**	0.41	**0.75**	0.53	**0.62**	0.17	0.59	0.27
labor	**0.67**	0.32	0.43	0.62	0.40	**0.49**	0.62	0.20	0.30	0.38	**0.68**	**0.49**
security	0.13	0.25	0.17	0.08	0.42	0.14	**0.15**	0.25	**0.19**	0.10	**0.92**	0.17
taxes	0.74	**0.82**	0.78	0.79	0.65	**0.71**	**0.80**	0.47	0.59	0.79	0.65	0.71
war_foreign	0.91	**0.79**	**0.85**	0.75	0.73	0.74	0.87	0.78	0.82	**0.94**	0.41	0.57

The results from the different models are represented in Table 5 and Fig. 5. As can be observed, gpt-4o yielded the results with highest accuracy, even though, close to those obtained with llama3-70b. In contrast, mDeBERTa yielded the lowest performance metrics, achieving an accuracy around 43%.

Table 5. Performance metrics for the implemented models.

performance metrics	gpt-4o	llama3-8b	llama-70b	mDeBERTa
accuracy	0.69	0.62	0.67	0.43
precision	0.74	0.71	0.73	0.62
recall	0.69	0.62	0.67	0.43
f1 score	0.70	0.63	0.68	0.46

The results of the Friedman test, with a statistic of 12.0 and a p-value of 0.00738, indicate a statistically significant difference between the models in terms of metrics (accuracy, precision, recall, f1 score). Since the p-value is less than 0.05, we reject the null hypothesis that all models have equal performance, suggesting that at least one model performs differently. The Nemenyi post-hoc test indicates that among the

analyzed models, only the performance of the gpt-4o model is significantly better than that of the mDeBERTa model, with a rank difference of 3.0, exceeding the critical difference (CD) of 2.345. Comparisons between the other pairs of models (gpt-4o vs. llama3-8b, gpt-4o vs. llama-70b, llama3-8b vs. llama-70b, llama3-8b vs. mDeBERTa, and llama-70b vs. mDeBERTa) did not show statistically significant differences. Therefore, it can be concluded that, according to this test, gpt-4o significantly outperforms mDeBERTa, while the differences between the other models are not statistically significant.

Table 6. Performance metrics for the implemented models (considering the match with any classification from the human raters).

performance metrics	gpt-4o	llama3-8b	llama-70b	mDeBERTa
accuracy	0.75	0.69	0.74	0.49
precision	0.80	0.78	0.80	0.68
recall	0.72	0.69	0.74	0.49
f1 score	0.76	0.70	0.75	0.51

Similar to what was seen previously, the results of the Friedman test, with a statistic of 11.15 and a p-value of 0.0109, indicate significant differences between the models. Also the Nemenyi test showed once again that only gpt-4o differs significantly from mDeBERTaModel 4, with a rating difference of 2.75, exceeding the critical difference of 2.345.

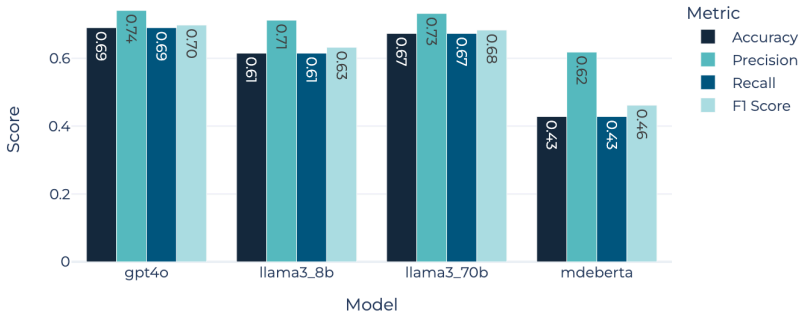

Fig. 5. Performance metrics for the implemented models.

4 Discussion

In this work, we implemented a pipeline for automatic speech recognition, speaker diarization and automated classification of political topics based on language models. The language models included mDeBERTa, a widely used model for NLP, including natural language understanding (NLU). Additionally, we assessed the performance of recently released LLMs, including gpt-4o and llama3 (8b and 70b versions). Overall, gpt-4o yielded the highest accuracy (69%) among the explored models, even though very close to the accuracy obtained with llama3-70b (67%). mDeBERTa yielded the less accurate predictions (43%) against manually-labeled classifications. The accuracy of the models improved further when considering secondary manual classifications, again, with very close results between gpt-4o (75%) and llama3-70b (74%). These results demonstrate the close proximity between open-source LLMs (namely those recently released by Meta) and SoTA closed-source LLMs. Also, the results demonstrate the relevance of automated pipelines for the extraction of knowledge from multimedia recordings of political debates.

It is worth mentioning that the elimination of speakers from the diarization output resulted from ad-hoc heuristics that were suitable for this context. These political debates followed a structure where moderators engage in a question-and-answer type of interaction with each of the political candidates. As such, knowing the number of candidates and moderators presented in each debate defined the ground rules for defining the number of speakers to be retained. The additional identified speakers corresponded to periods with superimposition between the intervention of a political candidate and, for instance, the reaction of the audience to what that candidate was saying. Also, in this research, there were non-relevant identifications for the purpose of political debates, pertaining to commercial activity exhibited during the breaks of the content.

While this research enabled an innovative classification of political topics during TV-broadcasted political debates, it also establishes the basis for deeper analysis, likely to produce valuable insights in the context. One potential approach may involve the use of complementary NLP approaches, such as sentiment analysis, which may play a pivotal role in identifying the positioning of political interventions during political debates, namely the expression of support, opposition, or neutrality on various issues. Finally, the proposed framework may also inspire research in other areas where transcription of multimedia recordings are frequently used for generating domain-specific insights. In communication research, this framework could support the study of media interactions, public speeches, and broadcast communications. The use of this approach can help researchers in dissecting communication patterns, identifying key speakers, and classifying topics with high accuracy. - ultimately, enabling a deeper understanding of how information is conveyed and received. Other relevant applications include psychological research, in which similar techniques can be employed to classify interactions during psychotherapy sessions or focus group discussions, or to summarize non-structured interviews.

5 Conclusions

This work demonstrates the viability of a pipeline involving fully automated automatic speech recognition, diarization and language models for the purpose of classifying political debates based on multimedia content.

6 Acknowledgments

The authors would like to acknowledge Bruno Ribeiro for his support on the definition of political topics. The authors would also like to acknowledge the support of Ana Luísa Abreu, Beatriz Caldeira and João Figueiredo for supporting this research with the manual classification of the political topics.

7 Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

8 References

1. Laver, M., Benoit, K., Garry, J.: Extracting Policy Positions from Political Texts Using Words as Data. Am. Polit. Sci. Rev. 97, (2003).
2. Doan, T.M., Baumgartner, D., Kille, B., Gulla, J.A.: Automatically Detecting Political Viewpoints in Norwegian Text. In: Miliou, I., Piatkowski, N., and Papapetrou, P. (eds.) Advances in Intelligent Data Analysis XXII. pp. 242–253. Springer Nature Switzerland, Cham (2024).
3. Lai, M., Celli, F., Ramponi, A., Tonelli, S., Bosco, C., Patti, V.: HaSpeeDe3 at EVALITA 2023: Overview of the Political and Religious Hate Speech Detection task. (2023).
4. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust Speech Recognition via Large-Scale Weak Supervision. (2023).
5. Plaquet, A., Bredin, H.: Powerset multi-class cross entropy loss for neural speaker diarization. In: INTERSPEECH 2023. pp. 3222–3226 (2023).
6. Bredin, H.: pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In: INTERSPEECH 2023. pp. 1983–1987. ISCA (2023).
7. Khoma, V., Khoma, Y., Brydinskyi, V., Konovalov, A.: Development of Supervised Speaker Diarization System Based on the PyAnnote Audio Processing Library. Sensors. 23, 2082 (2023).
8. Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder‐Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., Staab, S.: Bias in data‐driven artificial intelligence systems—An introductory survey. WIREs Data Min. Knowl. Discov. 10, e1356 (2020).
9. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced BERT with Disentangled Attention, (2021).
10. He, P., Gao, J., Chen, W.: DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, (2023).
11. OpenAI et al.: GPT-4 Technical Report, (2024).