

Argument-Mining from Podcasts Using ChatGPT

Mircea-Luchian Pojoni^{1,*}, Lorik Dumani¹ and Ralf Schenkel¹

Trier University, Behringstraße 13, D-54286 Trier, Germany

Abstract

Podcasts have emerged as a significant platform for the exchange of ideas, opinions, and knowledge on a variety of topics. At the same time, the extraction of arguments (called: argument mining) has received great attention. However, to the best of our knowledge, there exist no work that investigates the extraction of arguments from podcasts. One reason can be that podcasts often involve unpredictable and complex argument structures, and extracting valuable insights from them is challenging. In this work, we present the novel approach of extracting two different types of argumentative structures from podcast after transcribing them, i.e., (1) a simple but often used variant describing arguments as consisting of only a claim and a premise, where the claim describes the standpoint and the premise the reason to support or attack that claim and (2) an extended variant where an argument comprises premises, a main claim, counterarguments, and rebuttals. For this purpose, we utilize two specially designed prompts and OpenAI's GPT-4 language model. For our test data, we chose three podcasts considering current computational constraints and the need for diversity in topics and discussion styles. Our evaluation shows the high feasibility of extracting arguments from podcasts using ChatGPT. We publish the podcasts' transcripts as well as the extracted arguments.

Keywords

Argument Mining, Argument Graphs, ChatGPT, Podcasts, transcriptions

1. Introduction

Podcasts have become a popular and accessible way in which people share ideas, opinions and knowledge on a variety of topics. In addition to the convenience of gathering information even on the move, listeners particularly appreciate the authenticity of the speakers and conversations, as the monologues or dialogues are conveyed as they would be in real life, in contrast to written texts. It is also possible to retain knowledge of issues well when, as is often the case with podcasts, they are not communicated in a dry manner. That is, the language is not overly chosen as it is, for example, in essays, but it is also not too colloquial or even insulting as, for instance, in debate portals. Rather, it is as people would speak to each other in a real dialog. In addition, the arguments presented are often quick-witted, which is also in contrast to written texts.

In the field of computational argumentation, great progress has been achieved, especially in the field of argument mining (AM). In AM, the goal is to extract argumentative structures from unstructured natural language texts [1]. There are many definitions of what an argument is. Here, we will introduce and use two definitions:

ICCBR TMG'23: Workshop on Text Mining and Generation at ICCBR2023, July 17 – 20, 2023, Aberdeen, Scotland


*Corresponding author.

✉ s4mipojo@uni-trier.de (M. Pojoni); dumani@uni-trier.de (L. Dumani); schenkel@uni-trier.de (R. Schenkel)

🆔 0009-0003-2217-5864 (M. Pojoni); 0000-0001-9567-1699 (L. Dumani); 0000-0001-5379-5191 (R. Schenkel)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

A frequently used definition describes an argument consisting of a claim and a premise. The claim embodies a perspective that the transmitter aims to make more or less appealing to the receiver. For this purpose, the transmitter uses premises, which can serve as either oppositions or supporting evidence. An example of a claim would be “*nuclear energy has no future*”. A supporting premise to this claim is “*nuclear power plants are dangerous*”, an opposing premise to this claim is “*nuclear power is clean*”.

Another often used definition frames arguments as constituents of a larger, overarching topic or issue, often referred to as the *main claim* (or major claim) [2]. The advantage of using this framing lies in its implicit assumption of the inter-relatedness of arguments within an argumentative discourse. This inherently fosters a more organized placement of the arguments in relation to the discourse’s main claim, rather than treating them as disparate elements. As such, here we define the main claim as the core statement of the discourse and *premises* as statements which constitute implicitly supportive standpoints. *Counterarguments*, on the other hand, will be defined as statements challenging the truth value of the the main claim directly, or via challenging the truth value of the premises. Finally, we will also define and address certain statements as *rebuttals*, if they present themselves as a transmitter’s resolution between premise-counterargument pairs. For example, if the main claim is “*nuclear energy has no future*”, “*nuclear power plants are dangerous*’ would be a premise, and “*nuclear power is clean*” would be a counterargument, challenging the main claim directly (rather than challenging the truth-value of the premise). In this context, “*Despite the low environmental impact, the catastrophic consequences of a mishap, as shown by history, far outweigh the benefits of nuclear energy*” would constitute a rebuttal, because it is a statement that acknowledges both the aforementioned premise and counterargument, presenting a resolution by comparing their relative significance. It is worth noting that in this work we aim to identify and extract rebuttals as presented by the speaker, rather than rendering resolutions out of premise-counterargument pairs independently.

Overall, while a main-claim oriented strategy may avoid issues associated with reassessing the structural nuances of the discourse at a later stage, it may impose certain limitations on the breath of topics addressed. For instance, some arguments might touch upon a completely different overarching topic, which could potentially lead to a degree of content loss, despite the possibility that such content may still hold value overall. Considering the strengths and weaknesses of these approaches, in this paper, we will investigate AM using both definitions.

As mentioned before, there are already great advances in AM. However, to the best of our knowledge, these existing methods have so far only been applied to written texts such as essays, user-generated texts like debate portals, and debate/political speech derived text. Hence, in this paper, we investigate AM on podcasts, where the text has to be extracted first, of course. Specifically, the recent release of GPT-4 offers an extensive array of avenues and possibilities for furthering the scope of AM. With this work, we make the following contribution:

- 1 We publish a dataset consisting of approximately 1,500 podcast episode transcriptions. These episodes originate from 11 distinct shows that were selected due to their varied topics, discussion styles, and formats. A distinctive feature of this dataset is that elements such as timestamps and speaker identification labels are not present. This design decision underpins our belief that argument mining with podcast data hinges not primarily on metadata, but rather on the contextual interpretation of the content itself. By omitting

these identifiers, the dataset is streamlined to focus purely on the transcribed text, promoting research that prioritizes contextual understanding over structural dependencies. A subset of this dataset, including 71 episodes from 3 shows¹ has been utilized to investigate our AM approach. The remaining portion of the dataset² is provided to stimulate and facilitate further research beyond the scope of this project.

- 2 We utilized OpenAI's GPT-4, through the ChatGPT interface, to extract argumentative structures from each show in our test set. We have provided two base prompts, which align with the definitions we have previously discussed, as well as two datasets. The first dataset is divided into three segments - one for each show - comprising a total of 445 prompt-response pairs. These pairs were generated by sequentially prompting segmented sections from the respective episodes of each show. The second dataset is also divided into three parts and includes 60 manually assessed response samples to ensure optimal quality.³
- 3 We measured and analyzed the effectiveness of our approach in relation to concepts which we will introduce as "*handle accuracy*", "*stance quality*", "*semantic coverage*" and "*semantic coherence*". Finally, we will offer our assessment of the limitations, our final conclusions and potential areas for future research.

The remainder of the paper is organized as follows: Next, we discuss related work in Section 2. We then introduce the podcast dataset and how we obtained it in Section 3. Then, in Section 4, we present the methods for extracting the argumentative structures in these podcasts, followed by an evaluation in Section 5. In Section 6, we conclude the paper and provide an outlook on future work.

2. Foundations and Related Work

In this section, we survey important foundations and work related to our paper in extracting argument structures from podcasts with state-of-the-art methods. Hence, we give a brief overview of argument mining, explain approaches to transcribing podcasts, and briefly explain ChatGPT.

Argument Mining AM is an established research area in computational argumentation that is encompassed in natural language processing, knowledge representation and reasoning, and human-computer interaction. The goal of AM is to extract natural language arguments and their relationships from texts and provide machine-processable structured data for computational argumentation models. The main tasks in the AM framework are argument extraction and relationship prediction [1]. AM emerged around 2010, when the first methods for extracting arguments from natural language documents were proposed. Since then, the research area has seen rapid development [1]. More recently, speech data has gained importance in AM

¹These podcasts include Borrowed Future, On with Kara Swisher, and Politics Weekly America.

²All 1498 episode transcriptions can be downloaded in .txt format here: <https://zenodo.org/record/7985213>

³Both the base prompts and all investigative datasets can be found here: <https://zenodo.org/record/7988082>

through the use of political speech and debate speech, as demonstrated by Orbach et al. [3], who addresses detecting articles that effectively invalidate the arguments of a given text, focusing on text derived from debate speeches specifically. Despite great progress, there are still some open challenges in the field of argument mining. One such challenge is to develop approaches that generalize well to different text types and languages. This means that effective methods for argument mining are needed that can be successfully applied regardless of the specific text domain or language, taking into account their structural and semantic differences, without sacrificing their performance [1]. The development of such cross-domain and cross-language approaches poses a significant research challenge and requires close collaboration between experts from different disciplines. By addressing the use of podcasts in AM, we uncover a prolific source of data in terms of both nature and volume, and hope to provide an important keystone in addressing the challenges mentioned above.

Podcast Transcriptions for AM Podcasts are a valuable source of information and cover an enormous range of topics, from politics to health and economics. According to a 2021 study by Edison Research, an estimated 116 million Americans have listened to at least one podcast in the last month or so alone [4] which is more than one third of the population of the United States⁴. Apple Podcasts, the world's largest catalog of podcasts, includes over 1 million shows in more than 100 languages and 175 countries and regions (as of 2020) [5].

Despite the growing interest in podcasts and the abundance of available content, to the best of our knowledge there has been virtually no research on extracting arguments from podcasts. One possible reason is the difficulty of generating sufficiently accurate machine transcriptions of podcasts. These include the challenges posed by different languages, accents, the presence of multiple speakers and computational costs. In addition, discourse markers, interruptions and the overall unscripted and somewhat unstructured nature of podcasts may also play a significant role in making it difficult to extract arguments from transcriptions.

OpenAI's Whisper models [6] being designed as highly robust speech processing systems, are trained on a diverse array of audio from various environments, speakers, and languages. Despite a LibriSpeech clean-test Word-Error-Rate (WER) of 2.5, which is not quite considered state-of-the-art, Whisper outperforms on nearly all other datasets due to its unique robustness properties. Moreover, it is important to note that the LibriSpeech clean-test dataset, which resembles a single-speaker podcast in nearly ideal acoustic conditions, does not account for the often multi-speaker and varied acoustic scenarios typically encountered in real-world podcasts [7]. Given its impressive performance across different speakers, accents, and speech rates, we selected Whisper for transcribing our diverse podcast collection. It excels in terms of versatility, handling English and multilingual content (supporting a very wide array of languages) easily deployable, even on consumer hardware, at a reasonable computational cost [6].

Extracting argumentative structures from podcasts is challenging due to imperfect transcriptions and unique dialogue features like interruptions and informal language. Even advanced AI models like Whisper have difficulty with incoherent speech. Thus, a robust extraction method must account for this noise and by understanding context on multiple levels. Podcast transcriptions also differ structurally from other argumentative texts due to multiple speakers,

⁴<https://www.census.gov/popclock/>, accessed: 2023-05-03

repetitions, and playful dialogue styles. Additionally, multilingual podcasts present further challenges as different languages often require different analysis methods. However, recent advancements in Large Language Models (LLMs) may offer a context-aware and multilingual pathway.

LLMs and ChatGPT LLMs are a class of AI models designed to understand sophisticated instances of natural language and generate text with human-like accuracy. They are trained on vast quantities of textual data, leveraging transformer architectures and their self-attention mechanisms, first introduced in the groundbreaking work by Vaswani et al. [8]. This allows them to capture complex language patterns and generate coherent, context-aware text, which is useful for various natural language processing (NLP) tasks like text translation, summarization, and more. Besides BERT [9] one of the best-known LLMs is OpenAI's family of Generative Pre-trained Transformer (GPT) models [10]. A hallmark moment in the field of LLMs was OpenAI's introduction of InstructGPT, which involved techniques developed to fine-tuning GPT-3 and align it, as well as other LLMs, with user intent across a wide range of tasks [11]. This, in turn, lead to the development of what has become known as ChatGPT (or GPT3.5), a sibling model to InstructGPT, designed to interact with users in a conversational manner. It can answer follow-up questions, admit mistakes, challenge incorrect premises, and reject inappropriate requests. Trained using Reinforcement Learning from Human Feedback (RLHF), ChatGPT employs a similar methodology as InstructGPT, with slight differences in data collection setup. With reaching 100 million unique users within the first two months of its release, ChatGPT is reported to be the fastest-growing consumer internet app in history [12].

GPT-4 has demonstrated significant improvements over its predecessor, GPT-3.5 (ChatGPT base model). GPT-4 excels over previous language models, including GPT-3.5, on various NLP benchmarks, most notably, the Measuring and Maximizing Language Understanding (MMLU) benchmark and a simulated US bar exam. It exhibits superior performance in many languages, even low-resource ones like Latvian, Welsh, and Swahili [13]. Additionally, GPT-4 outperforms the Personally Identifiable Information (PII) detection tool Presidio in an experiment on the Text Anonymization Benchmark (TAB), even without any examples, indicating exceptional contextual understanding. This preliminary evidence highlights GPT-4's extensibility and potential for further improvement in context awareness [14]. Despite some of its limitations in other areas, and lack of application in argument mining at the time of this writing, GPT-4's multilingual capabilities and context-awareness make it a promising candidate for potential use in argument mining tasks. However, the key to successfully employing GPT-4 for argument mining lies in developing a proper rationale and carefully crafting instructions that guide the model to achieve the desired outcomes.

Prompt Engineering Prompt engineering involves designing effective prompts to guide large-scale language models, such as GPT-4, in generating accurate and contextually relevant responses. Basic prompts can be improved by providing more context or instructions, and formatting them as question-answering (QA) or few-shot prompting, which includes demonstrations or examples. A more complex prompt can consist of multiple elements, such as instructions, context, input data, and output indicators. Not all elements are required for every prompt,

and the format depends on the task at hand. Building on the understanding that complex prompts can include complex structures, various techniques have been developed to optimize the performance of language models. Introduced by Wei et al. in 2022, Chain-of-Thought (CoT) Prompting [15], for example, enables better reasoning performance via steering the model towards a intermediate-steps oriented approach. Yao et al.(2022) [16] developed ReAct Prompting, facilitating interaction with external tools and the generation of reasoning traces and task-specific actions. Lastly, "Tree of Thoughts" (ToT), introduced by Yao et al. (2023) [17], presents a novel framework that enhances language model's problem-solving abilities by allowing exploration over a coherent tree of text units, significantly improving performance on tasks requiring complex planning. By understanding these elements and techniques, users can optimize the performance of language models in various tasks.

In this work we apply zero-shot prompting with complex prompt structure. Accordingly, it is apparent that our methods can be improved with other prompting paradigms and LLM augmentation strategies. However, note that in this work, our goal is to draw attention to AM using podcast transcription data and to demonstrate, in a proof of concept fashion, that sufficiently capable LLMs can in principle be used to reliably extract arguments from podcasts. Thus, our work may serve as a baseline for later papers that may, for example, develop superior prompts.

3. Dataset

In this section, we present the dataset we obtained by transcribing podcasts via OpenAI's Whisper [6] and performing some pre-processing steps for the prompting process described in the next section.

3.1. Podcast Audio Data

The primary objective of this dataset is to establish a diverse and versatile foundation. Each podcast encompasses distinct topics, speaker counts, and formats, such as varying numbers of hosts, audio segments featuring speakers like politicians, short-form interviews, and long-form guest interviews. This diversity is crucial for enabling future research and testing the prompts under discussion. **Borrowed Future** focuses solely on the topic of student loans and provides in-depth discussions over its 464-minute length ⁵. **On with Kara Swisher**, spanning roughly 995 minutes, presents a broader scope covering technology, politics, and more, with interviews featuring prominent industry figures ⁶. Lastly, **Politics Weekly America**, with a length of about 1123 minutes, zeroes in on US politics. It includes diverse speakers and viewpoints, discussing a single topic per episode ⁷. It is worth noting that, at the time of research, two out of the three podcasts under examination, '*Politics Weekly America*' and '*On with Kara Swisher*', were still broadcasting, implying that the downloaded data may not encapsulate the full scope of these series. In contrast, the '*Borrowed Future*' podcast had concluded its run and hence, the downloaded content for this podcast represents the complete series. Table 1 presents a summary

⁵<https://podcasts.apple.com/us/podcast/borrowed-future/id1478188105>

⁶<https://podcasts.apple.com/tm/podcast/on-with-kara-swisher/id1643307527>

⁷<https://podcasts.apple.com/gb/podcast/politics-weekly-america/id1610749444>

Table 1

Summary of the differences between the three datasets. All podcasts involved multiple speakers.

Podcast	Topics	Files	# of Prompts	Minutes	Description
Borrowed Future	Student loans	10	28 (MC-PCR) / 54 (CP-S)	464	One topic over the entire podcast
On with Kara Swisher	Tech, politics and more	18	64 (MC-PCR) / 131 (CP-S)	995	Multiple topics per episode
Politics Weekly America	US politics	43	51 (MC-PCR) / 117 (CP-S)	1123	One topic per episode

of differences between these podcast datasets, addressing topics and subject matter, number of files and prompts prepared, as well as length and a brief description of the podcast’s structure.

3.2. Transcribing the Podcasts

In order to extract and store arguments from podcasts, we first need to put the podcasts into a text form, wherefore we transcribe them. As explained previously in Section 2, we use OpenAI’s Whisper [6] library for this. More precisely, we wrote a script that iterates over audio files and passes whole files to the models. Both .mp3 and .wav file formats are supported. There are different models, ranging from around 40 million parameters in size (e.g. `tiny`, `tiny.en`) to models with around 1.5 billion parameters (`large_model`) that can be used in transcription. Note that although smaller models offer an enhanced speed of processing, they are somewhat likely to compromise the quality of the final output, depending on the nature of the audio in question [6]. We sequentially tested all these models for the podcast transcriptions. After a manual inspection, we found that `small.en` provided the best quality-cost trade-off for our dataset.

Given the constraints on the input length allowed by the ChatGPT interface and the relatively extensive length of our two base prompts (that is, prompts excluding the actual podcast input), we needed to partition our transcript files. This segmentation process was fairly straightforward. With a maximum input length of ‘ y ’ (measured in characters), we divided the transcription of an episode file ‘ x ’ into ‘ z ’ files, without regard for the exact split point, so long as the length of each segment ‘ x_i ’ was less than ‘ y ’, and the length of ‘ $x_{i0}, x_{i1}, \dots, x_{iz}$ ’ was roughly equal. We remained neutral about the precise location of the split, because we aimed to assess the effectiveness of our approach without any interventions that might assist the model in identifying semantic boundaries. However, we were careful not to split the text in the middle of a word, as this could potentially cause unwanted changes to the overall semantic integrity of the text. The maximum length ‘ y ’ of a segment ‘ x_i ’ was contingent upon the kind of prompt it was incorporated into. Here we briefly introduce them as *CP-S* and *MC-PCR*.

Next, we will address the specifics of the formally mentioned base prompts, and the overall methodology in terms of prompting and response processing.

4. Methodology

Having completed the treatment of the transcript subset in the previous section, a subset that will be used in our forthcoming experiments, we are now prepared to elaborate our design methodology with respect to the base prompts. Furthermore, we will provide a detailed

exploration of our approach to prompting, complemented by an explanation of our technique for processing responses.

The formally mentioned prompts are labeled as CP-S, standing for “*Claim, Premise - Stance*”, and MC-PCR, representing “*Main Claim - Premise, Counterargument, Rebuttal*”. Each of these prompts embodies one of the two definitions of an argument we discussed in Section 1. For better understanding, please refer to the excerpt illustrated in Appendix. Both prompts lay out the structure of the desired output and are built upon concepts that we have termed as *argument unit* (AU), *handles*, and *argument graph unit* (AGU), although there are minor differences in the definitions between the prompts. Broadly speaking, an AU is a statement synthesized from a handle, which is a direct quotation from the text, and is intended to serve as an anchor for the text segment from which the AU has been derived. The term AGU is then employed in leveraging the former, as a higher layer of abstraction.

We would like to reiterate that the transcription files underwent no pre-processing, except for the segmentation into equal-length sections to conform to the maximum prompt length, as previously discussed. Initially, this podcast segment input limit was set at approximately 14,000 characters, a constraint we adhered to for MC-PCS. Subsequently, this limit was reduced to roughly 7,500 characters, which we applied to CP-S.

4.1. MC-PCR - Prompt Design

This particular prompt assigns the model with the task of pinpointing arguments that revolve around a single central claim. This encompasses the main claim itself, premises (which are inherently supportive in this context), counterarguments, and rebuttals within the given podcast text. In this case, an AGU is simply construed as a pair comprising a handle and its synthesized counterpart; a handle being a textual citation from the podcast text that serves as the signpost for the synthesized AU (also referred to as ‘synth’). An AU, conversely, could represent a main claim, a premise, a counterargument, or a rebuttal, while maintaining its exclusivity to its respective category. The structure of the MC-PCR tasks is as follows:

- Component definitions (AU, handle, AGU, stance)
- Rules for the output (AGU unique to its category)
- Structure of the output (hierarchical argument graph)
- Provide the podcast text (max 14.000 Characters)

Here, the model should present the hierarchical argument graph, starting with the main claim, followed by premises, counterarguments, and rebuttals.

4.2. CP-S - Prompt Design

With this prompt, the model is tasked to extract simple Argument Graph Units (AGUs) from the provided podcast text. Each AGU consists of a claim, a premise, and a stance (support or oppose) related to the claim. For each claim and premise, again, a text citation (handle) is provided from the podcast text to serve as the source for the synthesized argument unit (AU). The structure of the task is as follows:

- Definitions of the components (AU, handle, AGU).
- Rules for the output (unique AGU paring, stances).
- Structure of output (ordered list of AGUs)
- Providing the podcast text (max 7.5k chars)

The model should present the AGUs in an ordered list, providing at least one AGU with a supporting stance and one AGU with a rejecting stance. The list can be expanded to include additional AGUs, if possible.

4.3. Prompting and Response Processing

In our approach to prompting, we crafted a script that effectively functioned as a semi-automated dialogue manager. This script harnessed the qualities of GPT-4, interacting with it through a ChatGPT wrapper implementation in python⁸. Its core task was to parse a sequence of prompts from a few JSON files. These files were designed to contain both the prompt-response pairs and meta-information corresponding to the prompting process of each individual podcast. These prompts were essentially base prompts combined with the podcast segment input as detailed in the previous section. The script sequentially fed these prompts to the GPT-4 model, and saved the generated response as the second element of a tuple composed of the original, complete prompt and the model's response, i.e. the prompt-response pairs. That is, if the response passed our keyword test.

In both of our prompts, we defined a specific response structure and used a simple list of keywords to verify whether the response to a given prompt was consistent with our pre-defined structure. We employed the keywords "START" and "END" as markers to check if the response adhered to the structure and was completed. That is, if a response didn't begin with the keyword "START," we could reasonably deduce that it did not follow the specified structure. Similarly, if the final keyword was not "END," we could infer that the response either did not conform to the structure or was not completed for some reason. The aforementioned list of keywords also included a few prompt-specific ones, which further assisted us in determining if the individual structural prerequisites were fulfilled. For the MC-PCR prompt, these keywords were "Main claim," "Premise," "Counterargument," and "Rebuttal." For the CP-S prompt, the keywords were "AGU #1:," "Claim-AU," "Premise-AU," and "Stance." If a response didn't contain these specific keywords, the prompt would be reissued until it met our conditions. We found this strategy to be highly effective for our specific needs.⁹

⁸See <https://github.com/mmabrouk/chatgpt-wrapper> for more details.

⁹Our research was constrained by GPT-4 access limitations, allowing only 25 prompts every three hours in April of 2023, thus reducing our data mining capability.

5. Evaluation

5.1. CP-S

CP-S Evaluation The evaluation of CP-S consists of a multi-dimensional approach where different aspects of the response are measured systematically. We will attend to each metric one by one.

- **Handle Accuracy for Claims and Premises (HA)** : is a basic metric used to gauge the precision of the model's proficiency in interpreting and generating textual information. This is evaluated as a binary measure. A score of '1' suggests that the synthesized text aligns with the contextual argument that encompasses the quote in question, be it a claim or a premise. Conversely, a score of '0' indicates that the synthesized text doesn't align with the given context. To be more specific, the handle-synth pair is deemed accurate if the segment of text it underscores – typically about 2-3 sentences preceding and 2-3 sentences following the quote – is semantically in harmony with the synthetic text.
- **Premise Stance Quality (SQ)** : This metric evaluates the system's ability to accurately assess the stance of a premise. The rating is binary, with '1' indicating accurate stance generation and '0' indicating a miss.
- **Semantic Coverage (SC)** : This measures the model's ability to cover the full semantic scope and meaning of an input. It is evaluated on a scale of 1 to 5, where 1 indicates very poor semantic coverage (meaning the output does not adequately represent the input), 2 indicates insufficient semantic coverage, 3 indicates sufficient semantic coverage (meaning the output represents the key ideas of the input text, but not the full scope), 4 representing almost complete semantic coverage, and 5 signifies complete semantic coverage (meaning the output fully and accurately represents the input).

One expert in the field of argumentation, who is also one of the authors, examined a total of 30 responses, with 10 responses dedicated to each podcast under consideration. The subsequent findings are as follows:

CP-S Results Our CP-S based approach demonstrated good performance across all tested datasets. On average, it achieved near-perfect handle accuracy for both claims and premises at 0.992. This high score validates the robustness of the approach in generating contextually coherent text. Similarly, the model's mean average premise SQ score was strong at 0.864, demonstrating its adeptness at assessing nuanced textual relationships, and evaluating their stance. Although there was a slight dip in the '*Politics Weekly America*' dataset with a score of 0.75, this is still a decent result and indicative of the overall robustness. In terms of semantic coverage, our approach consistently performed well, achieving an average of 3.23. This signifies that the model regularly captured the key ideas from the input text, if not always fully capturing the complete semantic scope. Breaking it down by datasets, there are no significant outliers in any of the aspects we investigated.

In summary, the CP-S part of our approach showed very promising proficiency in context understanding, stance quality, and semantic coverage. The consistently high performance across

Table 2

Performance of CP-S showing the mean average of handle accuracy for claims and premises (HA), and stance quality of premises (SQ), as well as the mean semantic coverage (SC) for the three datasets.

Dataset	claim HA	premise HA	premise SQ	SC
<i>borrowed future</i>	1 (1,1,1,1,1,1,1,1,1,1)	1 (1,1,1,1,1,1,1,1,1,1)	0.892 (1,0.75,0.75,0.75,0.667,1,1,1,1,1)	3.4 (3,4,3,3,3,3,4,3,4,4)
<i>on with kara swisher</i>	0.975 (1,1,1,1,1,0.75,1,1,1,1)	0.975 (1,1,1,1,1,1,1,0.75,1)	0.95 (1,1,1,0.5,1,1,1,1,1,1)	3 (3,3,2,3,3,3,3,4,3,3)
<i>politics weekly america</i>	1 (1,1,1,1,1,1,1,1,1,1)	1 (1,1,1,1,1,1,1,1,1,1)	0.75 (0.667,1,0.8,0.75,1,1,0.75,0.75,1,1)	3.3 (3,4,3,3,3,2,3,4,4,4)
<i>average</i>	0.992	0.992	0.864	3.23

various datasets attests to the system’s robustness and reliability. However, the scores also suggest potential for further enhancements.

5.2. MC-PCR

MC-PCS Evaluation The MC-PCS evaluation extends the suite used for the CP-S to a broader range of argumentative components. This wider array provides a more comprehensive assessment of the models argumentation generation capabilities, especially with respect to coherence between these components. It is important to highlight that because stances are inherently conveyed, we do not evaluate stance quality as it is subsumed within the coherence measurement. Similarly, our analysis does not extend to assessing semantic coverage due to the intrinsic constraints of MC-PCR. This approach, by its very nature, is subject to certain limitations in terms of semantic coverage, chiefly stemming from the unpredictability introduced by our segmentation process. Consequently, our primary focus will be centered predominantly on evaluating semantic coherence.

- **Handle Accuracy for MC-PCR (HA):** Handle accuracy in the MC-PCS context maintains the same principle as in the CP-S evaluation, but extends to a more diverse set of argument components. Like in CP-S, a ‘1’ is assigned when the system’s synthesized text for an argument component is plausible, and a ‘0’ is assigned when it doesn’t meet this criterion. The difference here lies in the application of this measure to a wider array of argumentative elements: main claims, supporting premises, counterarguments, and rebuttals.
- **Semantic Coherence (SC) :** In the MC-PCS evaluation framework, “semantic coherence” is a crucial metric. It measures the logical and meaningful connections between the components of an argument: the main claim, premises, counterarguments, and rebuttals. The evaluation is binary - a ‘1’ is given when there’s a clear, logical, and meaningful relationship between the components being evaluated, and a ‘0’ when there is not. This framework ensures a comprehensive evaluation of the model’s ability to extract arguments that are not only coherent individually but also form a coherent, interconnected argument structure. We assess the semantic coherence between different pairs of argument components as follows:
 - **main claim ← premise** - If every premise is connected to the main claim, a ‘1’ is assigned. Otherwise, a ‘0’ is given.

Table 3

Performance of MC-PCR showing the mean average handle accuracies for the generated arguments for main claim, premises, counterargument, and rebuttal.

Dataset	premises	main claim	counterargument	rebuttal
<i>borrowed future</i>	0.975 (1,1,1,1,1,1,1,1,0.75,1)	0.9 (1,1,1,1,1,1,0,1,1,1)	1 (1,1,1,1,1,1)	1 (1,1,1,1,1,1)
<i>on with kara swisher</i>	0.967 (0.667,1,1,1,1,1,1,1)	0.9 (1,1,1,1,1,1,1,0,1)	0.95 (1,1,1,1,1,1,1,0.5,1,1)	1 (1,1,1,1,1,1,1,1)
<i>politics weekly america</i>	0.942 (1,0.75,1,1,1,1,0.667,1,1,1)	1 (1,1,1,1,1,1,1,1,1)	0.85 (1,0.5,1,1,1,1,1,0,1)	0.94 (1,1,1,1,0.5,1,1,1,1,0)
<i>average</i>	0.961	0.933	0.933	0.98

- **main claim** \leftarrow **counterargument** - If at least one counterargument is connected to the main claim, a '1' is assigned. If no such connection exists, a '0' is assigned.
- **premise** \leftarrow **counterargument** - If at least one counterargument is connected to at least one premise, a '1' is assigned. If there's no connection between any premise and counterargument, a '0' is given.
- **premise x counterargument** \leftarrow **rebuttal** - If every rebuttal is connected to at least one premise and one counterargument, a '1' is assigned. If any rebuttal is not connected to at least one premise and one counterargument, a '0' is assigned.

We again examined a total of 30 responses, with 10 responses dedicated to each podcast under consideration. The subsequent findings are as follows:

MC-PCR Results In order to measure the performance of MC-PCS, we again relied on several determinants, as we did previously with CP-S. This time, however, on different ones, because the two argumentative structures are different. Since we no longer have a simple division into claim and premise, but into main claim, premises, counterargument, and rebuttal, we first measured the handle accuracies, as we did before with CP-S. These are listed in Table 3. Although they are slightly lower than for CP-S, they are still highly accurate with average handle accuracies of 0.933 for the main claim, 0.961 for premises, 0.933 for counterarguments, and 0.98 for rebuttals. These high scores indicate the model's proficiency in identifying and synthesizing contextually coherent text segments for all components of an argument.

In terms of SC, we observe noticeable performance variations based on the podcast and type of connection. "On with Kara Swisher" somewhat deviates from the others, albeit not significantly. This is mostly due to errors in assessing the scope of a single main-claim in text segments that addressed a multitude of topics (e.g. casual conversations about politics with sudden shift towards discussing Elon Musk). Regardless of these fluctuations, the model's overall performance in identifying coherent connections between different arguments, as well as assessing their type, is remarkably high. Specifically, the average scores were 0.9 and 0.471 for the connections main claim \leftarrow premise, and main claim \leftarrow counterargument, respectively. Additionally, the connections premise \leftarrow counterargument, and premise x counterargument \leftarrow rebuttal, received commendable average scores of 0.9 and 0.863.

Furthermore, it is important to note that while a main claim \leftarrow counterargument score of 0.471 might appear low, signifying substandard performance, this is mostly due to our definition.

Table 4

Performance of MC-PCR showing the mean average values for the SC connections.

Dataset	main claim ← premise	main claim ← counterargument	premise ← counterargument	premise x counterargument ← rebuttal
<i>borrowed future</i>	1 (1,1,1,1,1,1,1,1,1,1)	0.714 (0,1,0,1,0,1,1,0,1,0)	1 (1,1,0,1,1,1,1,0,1,0)	1 (1,1,0,1,1,1,1,0,1,0)
<i>on with kara swisher</i>	0.7 (0,1,1,1,1,1,1,0,0,1)	0.3 (0,0,1,1,1,1,0,0,0,0)	0.7 (0,1,1,1,1,1,1,0,1,0)	0.889 (0,1,1,1,1,1,0,1,1,1)
<i>politics weekly america</i>	1 (1,1,1,1,1,1,1,1,1,1)	0.4 (1,0,0,1,0,0,0,1,1,0)	1 (1,1,1,1,1,1,1,1,1,1)	0.7 (1,1,-1,1,1,1,0,1,1,1)
<i>average</i>	0.9	0.471	0.9	0.863

In other words, counterarguments regularly attack the truth-value of the premises, rather than the main claim, simply because the premises, by nature, provide more angles for attacking. This is illustrated by the fact that almost all counterarguments fulfill the premise ← counterargument connection, if present.

Overall, the results achieved in this category are both surprising and exceptional. The high scores across varied podcasts and different connection types underscore the model’s impressive capabilities in handling complex argumentative structures, especially when the prompt rationale is more refined. This not only highlights the potential of leveraging modern AI technologies like ChatGPT for argument mining but also serves as a strong testament to their promise for further advancements in this field.

Limitations A notable limitation in our study was the variability in the model’s outputs, where identical prompts sometimes yielded different argument phrasing or framing. This inconsistency could pose challenges in contexts requiring reproducible outputs. Lack of access to the GPT-4 model’s API, and thereby the temperature parameter adjustment, restricted us from exploring potential mitigation strategies. Future studies with API access may offer insights into managing this variability.

6. Conclusion and Future Work

In this research, we leveraged state-of-the-art technology, specifically the GPT-4 model, to perform argument mining using a prolific, yet unrealized source of data - podcasts. This effort demonstrates the successful extraction of argument graphs from one-shot podcast dialogue transcriptions, marking a significant step forward, while also highlighting areas for potential growth. Our approach, centered on the use of advanced LLMs, allows for a deeper exploration of complex conversation dynamics present in podcasts, thereby presenting a promising path towards addressing the remaining open challenges in argument mining.

In light of our findings, we believe that podcast transcriptions warrant further scholarly exploration in the realm of computational argumentation. The nuances of spoken language and the unique characteristics of podcast dialogue create a complex but valuable resource for data analysis. Moreover, it is imperative that we devise more effective and nuanced prompt paradigms to better leverage the potential of large language models in processing such transcriptions. Looking ahead, an intriguing avenue for future research would be to apply a similar approach

to a wider range of linguistic inputs. Examining the behavior of large language models when exposed to diverse languages would provide valuable insights into the universal applicability and limitations of these models. This endeavor could significantly contribute to our understanding of cross-language text analysis and bolster the development of multilingual argument mining techniques.

Acknowledgments

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the projects ReCAP and ReCAP-II, Grant Number 375342983 - 2018-2024, as part of the Priority Program “Robust Argumentation Machines (RATIO)” (SPP-1999).

References

- [1] E. Cabrio, S. Villata, Five years of argument mining: a data-driven analysis, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), 2018.
- [2] L. Dumani, M. Bierter, A. Witry, A. Ludwig, M. Lenz, S. Ollinger, R. Bergmann, R. Schenkel, The recap corpus: A corpus of complex argument graphs on german education politics, in: 15th IEEE International Conference on Semantic Computing, ICSC 2021, Laguna Hills, CA, USA, January 27-29, 2021, IEEE, 2021. doi:10.1109/ICSC50631.2021.00083.
- [3] M. Orbach, Y. Bilu, A. Toledo, D. Lahav, M. Jacovi, R. Aharonov, N. Slonim, Out of the echo chamber: Detecting countering debate speeches, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
- [4] E. Research, The infinite dial 2021 (2021).
- [5] Apple inc., apple services now available in more countries around the world, 2020.
- [6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, 2022.
- [7] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An asr corpus based on public domain audio books, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2015).
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), 2017.
- [9] J. Devlin, M. Chang, K. Lee, 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019. doi:10.18653/v1/n19-1423.
- [10] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, in: Preprint, 2018.
- [11] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askill, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, arXiv preprint arXiv:2203.02155 (2022).
- [12] D. Milmo, agency, Chatgpt reaches 100 million users two months after launch, The Guardian (2023).
- [13] OpenAI, GPT-4 Technical Report, arXiv preprint arXiv:2303.08774 (2023).

- [14] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with gpt-4, arXiv preprint arXiv:2303.12712 (2023).
- [15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, arXiv preprint arXiv:2201.11903v6 (2022).
- [16] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, arXiv preprint arXiv:2210.03629 (2022).
- [17] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, arXiv preprint arXiv:2305.10601 (2023).

Appendix

.1. MC-PCR Prompting Example

The subsequent section presents the first of three prompt-response pairs drawn from the episode transcript 'RM6819947303.txt' of the podcast 'Borrowed Future.' For the sake of conciseness and clarity, the character count of the podcast text was truncated to 4,586 by omitting several segments. This is a representative sample for illustrative purposes; however, it's worth noting that the full-length input consists of 13,694 characters.

— Input prompt —

Extract the Argument Graph Units (AGUs) from this podcast text.

- An AGU is defined as: Argument Unit (AU) and unique handles.
- An AU is defined as: an AU can be a Main claim, premise (support for main claim), counterargument (attack on main claim), and rebuttal (challenging premise or counterargument). The AU is synthesized from the input text.
- A handle is defined as: string quote from the input text, which can be used to find and verify the text from which the AU was synthesized.

Building the Argument graph:

- The argument graph has only one Main claim.
- Each AU is unique to its category, and does not repeat.
- Try extracting AGUs even if the conversation is informal.

The Argument graph will be structured as follows:

START

- **Main claim AGU**
 - text (handle: "...")
- **Premise AGUs**
 - text (handle: "...")
 - ...
- **Counterargument AGUs**
 - text (handle: "...")
 - ...
- **Rebuttal AGUs**
 - text (handle: "...")
 - ...

END

podcast text:

Back in 1980, my high school guidance counselor called me down to his office to discuss my future. Based on some test scores, Mr. Dunbar wanted me to apply to James Madison or Penn State. I told him that money was an issue and that the community college down the road seemed like a better option (...) 26 bucks a credit, I could afford to experiment. So I did. Eventually, I earned an AA degree and started working. A year later, when I had saved some money, I transferred my credits to Towson University and with some help from my mom and dad, I got a bachelor's degree in communications. Total cost for all of it? Less than 10 grand. Point is, I was able to start working full time as soon as I graduated in my chosen field, free from the crushing weight of a student loan. **Now, we have one and a half trillion dollars in outstanding student loans** . thousands of college graduates unable to find work in their chosen field, thousands more who dropped out before graduating but still in debt, and millions of good jobs that nobody's trained to do or even excited about exploring. It's a disaster. I am concerned, sincerely, that we're pressuring teenagers to borrow vast sums of money in exchange (...) their potential. **Since I graduated, the cost of college has increased 1,120 percent** . Nothing so important has ever gotten so expensive so quickly. Not food, not energy, not real estate, not even health care. The question is why. **My liberal arts degree has served me really well** , and I would never discourage anybody who wants one to go for it, if they can afford it. From Ramsey Network, I'm George Camel, and this is Borrowed Future, a podcast (...) higher education. **In a world where going to college feels necessary to become a successful member of society** , both parents and students alike think it's the

next step. That it will give them a competitive advantage to get ahead. But if (...) entire system. One of the saddest parts about the student debt crisis is that **the people who least deserve it are the ones who are getting hurt the most**. People who went into debt to enter the middle class. People who didn't have super rich parents. People who chose jobs that weren't on Wall Street or in Hollywood. People who are (...) for failure. I think there's enough information in this world today that can tell you whether or not you truly need a four-year degree. So it comes down to this. **Does a degree absolutely have to happen for me to go into the field that I want to go into?** That is the ultimate qualifying question. So if not everyone needs a college degree, what type of education do they need? Here's billionaire, entrepreneur and investor Mark Cuban talking about his education experience. I mean, I'm in the technology industry now. I didn't do any technology when I was in high school or college. And I just happened to get a job and so I just took the time to use their computers to teach myself how to program. And that's how I learned everything. And I just dug in and said, this is it. It's make or break, right? I got to learn where I'm a toast. **If you're smart and you work hard, because you know what?** We live in an Internet age, right? You have access to everything you can take. If you want to know if you're MIT smart, you can go online and take an MIT class for free. Go prove you're smart. Walk into me when you're interviewing for a job and say, bam, look what I did with this class. I can only afford to go here. And I'm thinking, this person is smart and they're motivated and they're demonstrating to me that they know how to solve problems as opposed to I got all this debt and I saddled myself and now I have no choices. I just have to take the job that pays me the most. **You can go to college debt free, but you're going to have to do the work**. You're going to have to get that job. You're going to have to go to a college (...) other barbers who also went to trade school who also didn't rack up debt. Now he's going to get some of their fees, their chair fees. And now his six figures go to a little bit higher six figures, all because of a sixteen thousand dollar investment in trade school in nine months of his time. **Trade schools are not for everyone. But they are for a lot of people**. You want to be a carpenter, go to trade school. If you want to be a welder, go to a trade school. If you want to get creative, do the research. There are a lot of trade school programs out there that can teach you specifically for that field and get you equipped to go directly into that field as soon as you graduate. Now, here's the thing (...)

— Response —

START

- **Main claim AGU:**
 - The current education system and the pressure to get a college degree is causing a student loan crisis. (handle: " **Now, we have one and a half trillion dollars in outstanding student loans** ")
- **Premise AGUs:**
 - College costs have increased dramatically. (handle: " **Since I graduated, the cost of college has increased 1,120 percent** ")
 - People without wealthy parents suffer the most from student debt. (handle: " **the people who least deserve it are the ones who are getting hurt the most** ")
 - Not all careers require a four-year degree. (handle: " **Does a degree absolutely have to happen for me to go into the field that I want to go into?** ")
 - Trade schools provide an affordable alternative for some careers. (handle: " **Trade schools are not for everyone. But they are for a lot of people** ")
- **Counterargument AGUs:**
 - College degrees can still be valuable for some people. (handle: " **My liberal arts degree has served me really well** ")
 - Some believe that a college degree is necessary for success. (handle: " **In a world where going to college feels necessary to become a successful member of society** ")
- **Rebuttal AGUs:**
 - Self-learning and motivation can be a more effective path to success. (handle: " **If you're smart and you work hard, because you know what?** ")
 - There are alternative ways to get an education without accumulating debt. (handle: " **You can go to college debt free, but you're going to have to do the work** ")

END