Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Digital Engineering

Topic Segmentation using Large Language Models

Master's Thesis

Krishna Chaitanya Valaboju

1. Referee: Prof. Dr. Benno Stein
2. Referee: Jun.-Prof. Dr. Jan Ehlers

Submission date: March 14, 2025

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, March 14, 2025

.....
Krishna Chaitanya Valaboju

Abstract

Podcasts have emerged as a popular medium for delivering diverse content in informal and conversational formats. However, without effective topic segmentation, listeners struggle to navigate lengthy, unstructured transcripts, making it difficult to locate relevant discussions within an episode. The unstructured nature and inherent noise in podcast transcripts present significant challenges for topic segmentation. This thesis introduces two advanced segmentation methods: one that leverages large language models for semantic similarity thresholding with iterative refinement, and another that employs a transformer-based approach with BIO labeling for supervised sequence classification. Both methods are evaluated against a classical TextTiling baseline using manually annotated transcripts. The results demonstrate that incorporating deep semantic representations and contextual modeling leads to more accurate identification of topic boundaries, thereby enhancing content navigation and information retrieval in the podcast domain.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Dr.Johannes Kiesel, for his invaluable guidance, support, and constructive feedback throughout this research. I also thank my friends for their insightful suggestions and continuous encouragement, as well as my family for their unwavering support. This work would not have been possible without the inspiration and contributions of the entire academic community.

Chapter 1

Introduction

The exponential growth of podcasts as a medium for delivering news, entertainment, and educational content has created an unprecedented volume of spoken audio data. Unlike traditional written texts, podcast transcripts are characterized by informal language, conversational flow, disfluencies, and a lack of clear structural markers. These properties pose unique challenges for natural language processing, particularly for the task of topic segmentation, where the goal is to divide a long transcript into semantically coherent segments.

The motivation for this research stems from the observation that, as with YouTube videos, a significant portion of podcast content remains underutilized due to the absence of effective content organization. Without clear segmentation, users are forced to sift through lengthy, unstructured transcripts to locate relevant information. By automatically segmenting transcripts into topics, the accessibility of podcast content can be significantly improved. This not only enhances content navigation and summarization but also lays the foundation for downstream applications such as information retrieval and discourse analysis.

The task of segmenting podcast transcripts additionally shares conceptual similarities with video segmentation methods used for YouTube content (Figure 1.1). In YouTube video segmentation, the objective is to partition long videos into meaningful segments—often using a combination of visual, audio, and textual cues—to facilitate content navigation, summarization, and retrieval. Although podcast segmentation relies solely on the textual modality derived from automatic speech recognition, the underlying principles remain analogous. Both tasks require the identification of topic boundaries that enable efficient indexing and improved user experience. This work adapts and extends methodologies inspired by multimedia segmentation research, particularly those techniques employed in the context of YouTube videos, to the domain of podcast transcripts.

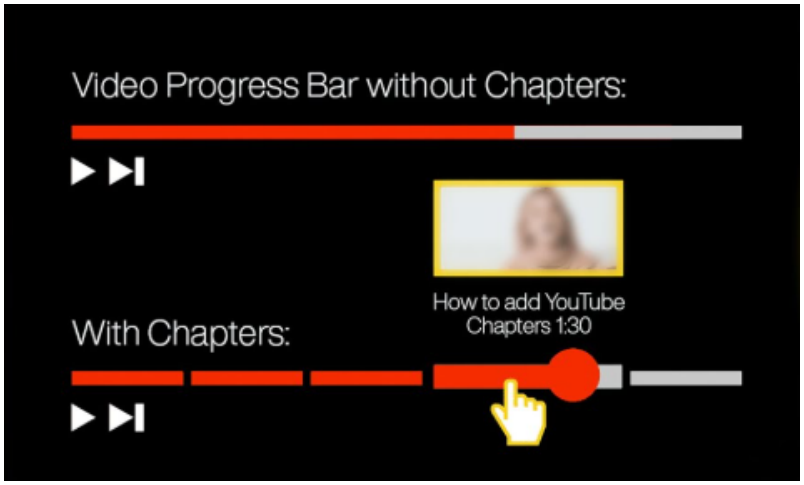

Figure 1.1: YouTube video segmentation with and without chapters, illustrating the difference in the progress bar. Source: `https://www.youtube.com/watch?v=WvPOshC74Og`

The primary goal of our research is to develop an automated system capable of decomposing lengthy podcast transcripts into meaningful segments that accurately reflect topic shifts within the discourse. To achieve this, we adopt a hybrid approach that integrates classical unsupervised methods with state-of-the-art deep learning techniques. We implement and evaluate three automated segmentation methods:

1. **TextTiling:** A classical algorithm that detects topic boundaries by analyzing lexical cohesion through sliding window comparisons. We capitalize on the distribution of words across text blocks to identify significant shifts in content.

2. **LLM-Based Topic Extraction with Similarity Thresholding:** We leverage the semantic understanding of large language models to extract primary topics from the transcript and assign sentences to topics based on the cosine similarity of their semantic embeddings. We generate these embeddings using models such as `all-mpnet-base-v2`, which capture nuanced semantic relationships between sentences.

3. **Transformer-Based BIO Labeling:** We employ a supervised method that uses a transformer model fine-tuned with the BIO (Begin, Inside, Outside) tagging scheme. This model segments the transcript into discrete topics by classifying each sentence, with our manually annotated corpus serving as the training data.

To rigorously assess our methods, we utilize a manually annotated corpus of podcast transcripts as the reference benchmark for evaluating the automated segmentation systems. Our annotation protocol adopts a hierarchical labeling scheme that distinguishes between main topics and subtopics while also accounting for transitional sentences. We conduct a quantitative evaluation using metrics such as the F1 score, P_k, and WindowDiff, ensuring that our system's performance is thoroughly measured against the gold standard. This multifaceted approach is designed to address the challenges posed by the noisy and variable nature of spoken content.

This thesis is organized as follows:

- **Chapter 2 – Background:** This chapter reviews the literature on topic segmentation in both written and spoken domains, highlighting the limitations of traditional approaches and motivating the need for advanced techniques tailored to podcast data.
- **Chapter 3 – Methodology:** Detailed descriptions of the data collection and preprocessing procedures, manual annotation protocol, and the implementation of the automated segmentation methods (TextTiling (Baseline), LLM-Based Topic Extraction, and Transformer-Based BIO Labeling) are provided.
- **Chapter 4 – Evaluation and Results:** The experimental design is outlined, including the selection of evaluation metrics and the benchmarking of automated segmentation methods against the manually annotated gold standard.
- **Chapter 5 – Conclusions:** A summary of the research contributions is presented, followed by reflections on the overall findings and suggestions for further investigations.

Chapter 2

Background

This chapter provides a foundational understanding of the key concepts, challenges, and advancements related to the topic segmentation of podcast transcripts using large language models (LLMs). The rapid growth of podcasting as a medium has resulted in an exponential increase in unstructured audio content, making it challenging for users to navigate and retrieve specific information effectively. Topic segmentation, a subfield of natural language processing (NLP), emerges as a critical tool in addressing this challenge by enabling the identification and separation of distinct themes within transcripts.

This chapter outlines the theoretical and practical aspects of topic segmentation, tracing its evolution from traditional methods to modern machine learning approaches. It also highlights the transformative role of LLMs in enhancing segmentation accuracy and efficiency. Furthermore, it discusses in detail the unique challenges posed by unstructured conversational data, such as podcast transcripts, and the specific methodologies developed to address these issues.

2.1 Introduction to Topic Segmentation

2.1.1 Definition and Importance

Topic segmentation is the process of partitioning a continuous stream of text or transcript into discrete, coherent segments, where each segment represents a distinct topic or subtopic. This fundamental NLP task facilitates numerous downstream applications such as information retrieval, text summarisation, and content indexing by transforming unstructured and often lengthy texts into manageable, semantically meaningful units.

In traditional settings, early methods like TextTiling Hearst (1997) relied on detecting shifts in lexical cohesion. TextTiling operates by dividing text

into blocks and analysing the distribution and co-occurrence of words to infer topic boundaries. While this approach proved effective for well-structured texts such as academic articles and news reports as it is less adept at handling the informal, unstructured, and conversational nature of spoken-word content.

Podcast transcripts exemplify this challenge. They are characterised by spontaneous speech, lack explicit formatting, and are often marred by errors introduced during automatic speech recognition (ASR). Consequently, accurately detecting topic boundaries in such data requires methods that go beyond surface-level lexical statistics. Recent advancements have harnessed the power of large language models (LLMs) to address these issues. For example, the PODTILE model Ghazimatin et al. (2024) leverages a transformer-based architecture (using models such as LongT5) to jointly generate chapter boundaries and descriptive titles for podcast transcripts. This model incorporates both static context such as episode metadata and dynamic context-like previously generated chapter information to better capture long-range dependencies and subtle semantic shifts in conversational speech.

The evolution from rule-based methods such as TextTiling Hearst (1997) to LLM-based approaches represents a significant leap forward in topic segmentation. Modern transformer-based models Vaswani et al. (2023) are not only capable of detecting abrupt topic shifts but also excel in identifying nuanced transitions where speakers blend topics Ghazimatin et al. (2024). This enhanced capability is critical for improving the user experience on podcast platforms, where clear content organization can dramatically enhance navigation and retrieval. By enabling functionalities such as chapter-based navigation and targeted search within episodes, advanced topic segmentation techniques facilitate quicker content understanding and more efficient information retrieval (Devlin et al., 2019; Liu et al., 2019).

Thus, topic segmentation plays a vital role in structuring unstructured audio content, thereby improving the accessibility and usability of large-scale, conversational data. Its importance is underscored by both the foundational traditional approaches and the modern, LLM-powered solutions that continue to push the boundaries of what is achievable in noisy, unstructured domains.

2.1.2 Applications of Topic Segmentation

Topic segmentation plays a pivotal role in several domains, enabling a wide array of applications that directly or indirectly enhance the usability of textual data. Some key applications include:

1. **Information Retrieval:**
   Topic segmentation plays a pivotal role in information retrieval by enabling the indexing of documents at a finer granularity. When long-form

content such as a podcast transcript is divided into coherent segments, each segment can be treated as an individual unit for retrieval purposes. This allows search engines to match queries not to an entire lengthy transcript but rather to specific, relevant segments. Early approaches like TextTiling, introduced by Hearst (1997), demonstrated that shifts in lexical cohesion could signal topic boundaries, thereby offering a mechanism to segment texts effectively. More recent methods, including probabilistic Bayesian models Eisenstein and Barzilay (2008) and transformer-based systems such as PODTILE Ghazimatin et al. (2024), have further refined this process. These modern approaches capture subtle semantic shifts by leveraging deep contextual embeddings, which enhance the precision of segment retrieval. In practical applications, users searching for a specific subject within a podcast can be directly routed to the most relevant chapter, thereby reducing retrieval noise and improving the overall search experience.

2. **Summarization:**
   Segmenting a document into discrete topics naturally lends itself to effective summarisation. By isolating distinct thematic units within a text, summarisation models can generate concise and coherent summaries for each segment. This process is particularly beneficial for long-form audio content such as podcasts, where users may only be interested in a brief overview of a specific discussion point. Recent studies have integrated segmentation with summarisation models using architectures like T5, BART, and Pegasus to automatically generate segment-specific summaries that closely mirror human-curated chapter titles Aquilina et al. (2023). These segment-level summaries not only aid users in quickly understanding the content but also reduce cognitive load by distilling large amounts of information into accessible highlights. (Joty et al., 2013).

3. **Content Navigation:**
   One of the user-friendly applications of topic segmentation is enhanced content navigation. In multimedia platforms, particularly those hosting podcasts, segmented content can be transformed into interactive navigation aids such as clickable chapter markers or timestamped indexes. For example, by applying segmentation algorithms to a podcast transcript, a platform can automatically generate chapters that users can click on to jump directly to the section of interest. The PODTILE model exemplifies this approach by integrating episode metadata with dynamic segmentation outputs, thereby providing clear, organized chapters that facilitate quick navigation (Ghazimatin et al., 2024). This method significantly improves user engagement and overall experience by reducing the

time and effort required to locate specific content within lengthy audio files.

4. **Recommendation Systems:**
   The ability to discern and isolate topics within content enhances the effectiveness of recommendation systems. When podcast transcripts are segmented into discrete topical units, recommendation algorithms can analyse these segments to better understand the content's thematic structure. This fine-grained analysis allows for more personalized recommendations, as the system can match user preferences not just to an entire episode but to specific segments that align with their interests. For instance, a listener interested in technology might be directed to segments that discuss recent advancements in artificial intelligence rather than receiving a generic podcast recommendation. Multimodal segmentation approaches, which combine text and audio embeddings Ghinassi et al. (2023), further refine these recommendations by capturing both semantic and acoustic features that contribute to a richer understanding of content.

5. **Knowledge Management:**
   In both academic and corporate environments, effective knowledge management hinges on the ability to structure and retrieve vast amounts of unstructured data. Topic segmentation facilitates this by organizing content into well-defined, searchable segments. For instance, segmented podcast transcripts can be archived and indexed according to topic, enabling quick access during research, training, or decision-making processes. This structured approach not only enhances searchability but also allows organizations to maintain a coherent repository of knowledge that can be easily referenced and cross-analyzed. Studies like those by Kazantseva and Szpakowicz (2012) emphasize the importance of segmentation in managing and utilizing large datasets, thereby highlighting its value for systematic information management.

6. **Sentiment and Emotion Analysis:**
   Segment-level analysis is critical for conducting detailed sentiment and emotion analysis in conversational data. When a podcast transcript is segmented by topic, each segment can be individually analyzed to determine shifts in sentiment or emotional tone. This granular approach allows researchers to pinpoint which parts of a conversation elicit positive, negative, or neutral responses, offering insights into audience engagement and the emotional impact of specific topics. For example, by isolating segments where speakers express enthusiasm or frustration, analysts can

better understand listener reactions and adjust content recommendations accordingly. Recent work on sentiment analysis in conversational data supports this approach by demonstrating that segmentation enhances the accuracy of emotion detection within complex, multi-speaker dialogues (Yu et al., 2023).

These applications underscore the versatility and importance of topic segmentation in managing and utilizing vast quantities of unstructured text, particularly in the growing domain of podcast transcripts. The next section will delve into the challenges that make topic segmentation, especially in the context of podcasts, a complex and evolving research area.

2.1.3 Unique Characteristics of Podcast Transcripts

Podcast transcripts differ fundamentally from traditional written texts due to several distinctive characteristics:

- **Conversational Dynamics:** Unlike formal documents, podcast transcripts capture spontaneous, unedited speech that often includes filler words, hesitations, and interruptions. These features lead to frequent topic digressions and overlapping content, complicating the clear demarcation of topic boundaries (Kazantseva and Szpakowicz, 2012).

- **Lack of Structural Cues:** Podcasts generally lack explicit formatting markers such as paragraphs, headings, or section breaks. This absence of structure makes it challenging for traditional segmentation algorithms, which typically rely on such cues to identify topic shifts.

- **Variability in Vocabulary:** Conversational language in podcasts tends to be informal and dynamic. Speakers may use colloquial expressions and domain-specific terms or vary their vocabulary rapidly throughout a conversation, which further complicates the detection of semantic continuity.

Understanding these unique characteristics is critical, as they directly impact the design and effectiveness of topic segmentation methods tailored for podcast transcripts.

2.1.4 Data Quality and Preprocessing Challenges

The quality of podcast transcript data significantly impacts segmentation performance. Several factors contribute to the challenge:

- **ASR-Induced Noise:** Automatic Speech Recognition (ASR) systems, while increasingly accurate, still produce transcripts with errors. According to Song et al. (2022), issues such as misrecognized words, incorrect punctuation, and omissions are common. These inaccuracies introduce noise, making it difficult for segmentation algorithms to detect true topic boundaries reliably.

- **Inconsistent Formatting:** Transcripts generated from podcasts often lack consistent formatting. Segmentation algorithms struggle to apply traditional rules effectively without clear paragraph breaks or punctuation. Minimal preprocessing is typically employed to preserve the natural flow of conversation. However, this also means that many textual artefacts remain, complicating the segmentation task.

- **Preprocessing Trade-offs:** There is a balance to be struck between cleaning the data and preserving the conversational context. Aggressive preprocessing (e.g., removing filler words) might eliminate useful signals about the speaker's intent, while too little preprocessing leaves the data noisy. Beltagy et al. (2020) discuss similar challenges when processing long-form texts, emphasizing the need for techniques that maintain the integrity of the original content while reducing noise.

Addressing these data quality issues is essential for developing robust segmentation models, as even state-of-the-art methods can be significantly affected by the inherent noise in podcast transcripts.

2.1.5 Challenges in Topic Segmentation

Topic segmentation, while immensely useful, comes with its challenges, especially when dealing with unstructured and conversational data like podcast transcripts. Below are the key challenges:

1. **Lack of Clear Topic Boundaries:**
   Podcasts often exhibit informal and conversational speech, leading to ambiguous or overlapping topic boundaries. Speakers may shift topics abruptly without explicit cues, or intertwine multiple themes within a single discussion. This complexity poses significant challenges for segmentation models aiming to delineate distinct topics(Gklezakos et al., 2024).

2. **Data Noise:**
   Transcripts generated via Automatic Speech Recognition (ASR) systems frequently contain errors due to factors like accents, background noise, or overlapping speech. These inaccuracies can mislead segmentation algorithms, diminishing their effectiveness. Addressing ASR errors is crucial for accurate topic segmentation (Song et al., 2022).

3. **Conversational Nature of Podcasts:**
   The dynamic and informal nature of podcast conversations, often involving multiple speakers with varied styles and vocabularies, adds complexity to segmentation tasks. Traditional models, typically designed for structured texts, may struggle with such unstructured data. Developing models that can handle conversational dynamics is essential (Ghinassi et al., 2023).

4. **Lack of Labeled Data:**
   The scarcity of annotated podcast datasets for training and evaluation hampers the development of robust segmentation models. Unlike domains with standardized benchmarks, podcast segmentation lacks extensive labeled data, complicating progress. Creating and sharing annotated datasets is vital for advancing this field (Kazantseva and Szpakowicz, 2012).

5. **Long-form Content:**
   Podcasts often span extensive durations, posing challenges for models with fixed context windows. Processing such long-form content requires hierarchical or chunk-based approaches to manage the data effectively. Innovative methods are needed to handle the length and complexity of podcast transcripts (Beltagy et al., 2020) (Gklezakos et al., 2024)

Addressing these challenges necessitates innovative methods, such as leveraging advanced Large Language Models (LLMs), multimodal approaches, and developing robust pre-processing techniques to handle noisy and unstructured data. The subsequent sections will explore the evolution of topic segmentation techniques and the role of LLMs in addressing these challenges.

2.2 Evolution of Topic Segmentation Techniques

2.2.1 Traditional Approaches

Over the past decades, topic segmentation has evolved from simple rule-based methods to sophisticated deep learning architectures. Early methods primarily

relied on surface-level lexical features, while modern approaches leverage powerful transformer-based models to capture both local and global context. In this section, we review the progression from traditional techniques to emerging hybrid models and discuss how evaluation methods have been refined alongside these advances.

TextTiling and Early Innovations

A seminal example of a traditional approach is the TextTiling algorithm introduced by Hearst (1997). TextTiling detects topic boundaries by analyzing changes in word frequency distributions and lexical patterns across a text. By segmenting text into blocks and calculating lexical similarity between adjacent blocks, it identifies likely topic shifts. This method proved effective for structured texts, such as scientific articles and formal documents, where topics change explicitly and predictably. However, it faced challenges in handling unstructured or conversational data, where topic boundaries are less defined or overlap.

Limitations of Rule-Based Methods

While methods like TextTiling Hearst (1997) offered an efficient way to identify topics by analyzing local changes in lexical cohesion, they relied heavily on fixed rules and statistical patterns that often fail to capture the complexities of human language. In particular, these algorithms assume that sudden shifts in word usage or frequency distributions reliably signal topic boundaries. Although effective in structured, monologic texts (e.g., academic articles), this assumption does not hold up well in conversational settings such as podcast transcripts where topic shifts can be abrupt, implicit, and intertwined with speaker-specific language styles (Kazantseva and Szpakowicz, 2012).

1. **Semantic Relationships:**
   Traditional rule-based methods typically depend on surface-level lexical overlap and frequency statistics. As a result, they overlook deeper semantic connections between sentences or paragraphs, such as paraphrasing, synonymy, and implicit references. For instance, two segments discussing the same topic but using different vocabulary would appear unrelated to a purely rule-based system Choi (2000). Consequently, these methods struggle with the rich semantic variations in spontaneous speech, limiting their ability to accurately delineate topic boundaries.

2. **Context Awareness:**
   Rule-based approaches generally operate on localized text windows using

features like word distributions in fixed-length chunks or adjacent paragraphs and thus cannot model long-range dependencies Hearst (1997). This limitation is especially problematic for podcasts, where a single thematic discussion may extend across multiple segments and weave in and out of subtopics Joty et al. (2013). Without a mechanism to retain contextual information beyond the immediate vicinity, these methods fail to capture overarching thematic continuity, resulting in segmentation that may be too granular or misaligned with actual topic transitions.

3. **Dynamic Vocabulary:**
   In conversational domains like podcast transcripts, speakers often employ colloquial language, domain-specific terminology, filler words, or abrupt changes in style Song et al. (2022). Rule-based methods relying on predetermined lexical cues cannot easily adapt to these variations. Their sensitivity to out-of-vocabulary words or rare terms means that even slight shifts in wording or style can lead to false positives or missed boundaries. Consequently, the "one-size-fits-all" nature of these rule-based algorithms proves insufficient for the diverse and evolving vocabulary found in real-world podcasting scenarios.

Advances Inspired by TextTiling

TextTiling inspired several subsequent methods, including C99 by Choi (2000), which introduced clustering algorithms for topic segmentation. This method used sentence similarity matrices and clustering to infer boundaries, offering improvements over TextTiling by incorporating more flexible measures of lexical similarity. However, like its predecessor, it was constrained by its reliance on surface-level features.

Summary of Traditional Approaches

Traditional approaches like TextTiling and C99 laid the groundwork for topic segmentation by addressing the need for automated topic detection. Their innovations in lexical cohesion analysis provided a foundation for later methods but highlighted the need for more sophisticated models capable of understanding semantic relationships and contextual dependencies. As the limitations of these methods became apparent, researchers turned to probabilistic and machine learning-based approaches, marking the next stage in the evolution of topic segmentation. Notable among these is the TextTiling algorithm by Hearst (1997), which segments text based on shifts in word frequency distributions. For example, TextTiling divides a document into fixed-size blocks and computes the lexical similarity between adjacent blocks using cosine similarity.

When the similarity score between blocks drops significantly, it is considered a potential topic boundary. This approach works well for identifying clear shifts in topics, such as those in structured text like academic articles or news reports. These methods, while effective for structured and formal text, often failed when applied to conversational or unstructured data due to their reliance on surface-level indicators rather than deeper semantic understanding. For instance, in podcast transcripts, where conversations often involve overlapping speakers and shifts in tone or style, traditional methods struggle to recognize implicit topic transitions. Similarly, chat-based data, such as asynchronous team communications or informal social media threads, presents challenges due to the fragmented and non-linear nature of interactions, which are difficult to segment accurately using rule-based approaches.

2.2.2 Probabilistic Models

The introduction of probabilistic models marked a pivotal moment in the evolution of topic segmentation techniques. Unlike rule-based approaches, probabilistic models are grounded in statistical principles, allowing for greater flexibility and adaptability to diverse text structures. These models view topic boundaries as latent variables inferred through probabilistic reasoning, leveraging statistical dependencies within the text to identify changes in topics.

Latent Dirichlet Allocation (LDA)

One of the most influential probabilistic models is Latent Dirichlet Allocation (LDA), introduced by Blei et al. (2003b). LDA models a document as a mixture of latent topics, each represented by a probability distribution over words. The model assumes that each word in the document is generated by one of these topics, which are inferred through a generative process.

Although LDA is primarily a topic modeling algorithm rather than a topic segmentation model, it has been adapted for segmentation tasks by applying it sequentially across segments of text. In such applications, LDA identifies shifts in dominant topic distributions, which may serve as potential topic boundaries. However, this is an indirect use case rather than an inherent functionality of LDA.

For instance, in processing long documents or transcripts, LDA can identify areas where the probability of a particular topic sharply decreases, which may indicate a topic shift. While this approach excels at detecting global topics and thematic structures, it struggles with fine-grained segmentation tasks, where local coherence and sentence-level dependencies play a critical role (Griffiths and Steyvers, 2004)(Blei et al., 2003b). Consequently, LDA is often used in

combination with other segmentation techniques to improve the accuracy of topic boundary detection.

Hidden Markov Models (HMMs)

Hidden Markov Models (HMMs) represent another cornerstone of probabilistic topic segmentation. HMMs treat text as a sequence of observations generated by hidden states, each corresponding to a topic. The transitions between states are governed by probabilities learned from the data, allowing HMMs to explicitly model the dynamics of topic transitions.

(Beeferman et al., 1999) were among the first to apply HMMs to topic segmentation, showing that their probabilistic nature could better capture the inherent variability in topic transitions than deterministic methods. By modeling each sentence or paragraph as an observation, HMMs infer the most likely sequence of topic states. Despite their advantages, HMMs face limitations in capturing long-range dependencies and contextual nuances, as they rely on a fixed number of states and predefined transition probabilities.

Bayesian Models

Bayesian models extend probabilistic methods by incorporating prior knowledge about text structure into the segmentation process. For example, the Bayesian Segmentation Model proposed by Eisenstein and Barzilay (2008) introduced priors on linguistic features such as discourse markers and semantic coherence. Bayesian models aim to maximize the posterior probability of a segmentation given the observed text and prior knowledge, making them more robust to noise and adaptable to diverse domains.

Eisenstein and Barzilay's work demonstrated the efficacy of Bayesian models in segmenting multi-party dialogues, where traditional methods struggled. By leveraging linguistic cues, such as speaker turns and dialogue acts, their approach improved segmentation accuracy in noisy and conversational datasets.

Extensions to Probabilistic Models

Over time, researchers sought to enhance probabilistic models by combining them with other techniques. Hybrid approaches, such as combining LDA with Markov models, aimed to integrate the strengths of topic modeling and sequential state transitions Minkov and Cohen (2007). Additionally, variational methods were introduced to improve the scalability of Bayesian inference, enabling probabilistic models to handle larger datasets and more complex text structures (Blei et al., 2003a).

Limitations of Probabilistic Models

Despite their contributions, probabilistic models exhibit several limitations:

1. **Lack of Contextual Awareness:**
   Probabilistic models like LDA and HMMs often fail to capture long-range dependencies and contextual relationships essential for fine-grained segmentation.

2. **Computational Complexity:**
   Inference in these models can be computationally expensive, particularly for large datasets or long-form text.

3. **Dependence on Assumptions:**
   Probabilistic models rely on assumptions about the generative process of text, which may not align with real-world text structures, especially in unstructured or conversational data like podcast transcripts.

Contributions to Future Models

While probabilistic models are no longer the state-of-the-art for topic segmentation, their principles continue to influence modern techniques. By framing topic segmentation as a probabilistic inference problem, these models laid the groundwork for machine learning and neural network-based approaches. For example, probabilistic reasoning remains integral to transformer-based models, where attention mechanisms compute weighted probabilities over contextual information.

Probabilistic models represent a significant step in the evolution of topic segmentation. Their focus on statistical dependencies and generative processes provided a foundation for later methods that sought to address their limitations through more advanced computational techniques.

With the advent of deep learning, neural network-based models became prominent in NLP. Recurrent Neural Networks (RNNs) and attention-based mechanisms improved the capacity to capture long-range dependencies in text. Joty et al. (2013) demonstrated the use of neural models for segmenting asynchronous conversations, leveraging features such as dialogue coherence and syntactic patterns. However, these models were constrained by their dependency on large annotated datasets, which are often scarce for domains like podcasts.

2.2.3 Transformer-Based Models

The introduction of transformer architectures marked a paradigm shift in natural language processing (NLP), including topic segmentation. Unlike previous models, transformers use self-attention mechanisms to capture dependencies across entire sequences, enabling a deeper understanding of context and relationships between words. These models offer significant advantages in both structured and unstructured text segmentation.

The Transformer Architecture

Introduced by Vaswani et al. (2023), the transformer architecture was designed to overcome the limitations of recurrent neural networks (RNNs) and long short-term memory networks (LSTMs). Key features of transformers include:

1. **Self-Attention Mechanism:**
    - The self-attention mechanism enables the model to weigh the importance of different words in a sequence, regardless of their distance from each other.
    - This allows transformers to capture long-range dependencies, which are crucial for identifying topic boundaries in lengthy or conversational texts.
2. **Parallel Processing:**
    - Unlike RNNs, which process text sequentially, transformers process entire sequences in parallel, significantly improving computational efficiency.
3. **Positional Encoding:**
    - Transformers incorporate positional encodings to retain the order of words in a sequence, ensuring that syntactic and semantic relationships are preserved.

BERT and RoBERTa

BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2019) was one of the first transformer-based models to achieve state-of-the-art performance in multiple NLP tasks, including topic segmentation. BERT's bidirectional training allows it to consider both left and right contexts, making it highly effective for understanding nuanced topic shifts.

RoBERTa (Robustly Optimized BERT Approach), introduced by Liu et al. (2019), improved upon BERT by optimizing pretraining techniques, including training on larger datasets and using dynamic masking. These enhancements made RoBERTa more robust for segmenting long-form text, such as podcast transcripts.

For topic segmentation, BERT and RoBERTa have been adapted to:

- **Detect Topic Boundaries**: By training the model to predict whether adjacent sentences belong to the same topic.
- **Classify Segments**: Assigning labels to segments based on their dominant topic.

GPT and Transformer Decoders

Generative transformer models such as GPT have been applied to topic segmentation tasks, offering a complementary approach to discriminative models like BERT. Unlike BERT which is bidirectional and optimized for understanding context from both the left and right of a token–GPT is trained in an autoregressive, unidirectional manner. This means that GPT generates text by predicting the next word based solely on the preceding context, which inherently promotes the creation of coherent and contextually consistent sequences (Vaswani et al., 2023)(Brown et al., 2020)

For the purpose of topic segmentation, this generative capability can be leveraged in several key ways:

- **Contextually Coherent Summarization:** GPT can generate concise summaries for individual segments of text. By summarizing content in a manner that preserves the essential meaning, GPT provides an implicit signal regarding the main topics discussed within that segment. When applied across a transcript, abrupt changes in the generated summaries can indicate potential boundaries between distinct topics. This approach has proven effective in identifying topic shifts, particularly in unstructured or conversational texts such as podcast transcripts (Yenduri et al., 2023).
- **Boundary Reinforcement through Latent Topic Extraction:** In addition to summarization, GPT's generative process can extract latent thematic cues that might not be evident through surface-level lexical matching. By producing output that reflects deeper semantic content, GPT supports the detection of subtle transitions that traditional rule-based or statistical methods might overlook.

Overall, while GPT's unidirectional nature means it does not directly model the full bidirectional context as BERT does, its strength lies in generating extended, coherent narratives. This quality makes it particularly effective in scenarios where maintaining a smooth, logically connected summary is crucial for accurately identifying topic boundaries in lengthy, conversational texts.

BERTopic

Introduced by Grootendorst (2022), **BERTopic** combines transformer-based embeddings with clustering algorithms to identify topics within document collections. This approach generates document embeddings using pre-trained transformer models, clusters these embeddings, and employs a class-based TF-IDF procedure to extract coherent topic representations. BERTopic has demonstrated effectiveness across various benchmarks, offering a robust method for topic modeling.

Transformer² Framework

Lo et al. (2021) proposed the **Transformer² framework**, which integrates pre-trained transformers as sentence encoders with an upper-level transformer-based segmentation model. This architecture captures semantic coherence within text segments, leading to improved performance in text segmentation tasks. The model benefits from both single and pair-wise pre-trained knowledge, enhancing its ability to detect topic boundaries.

Probabilistic Topic Modelling with Transformer Representations

Reuter et al. (2024) introduced the **Transformer-Representation Neural Topic Model (TNTM)**, which combines the strengths of transformer-based embedding spaces with probabilistic modeling. This model leverages the contextual representations provided by transformers to improve the coherence and interpretability of the extracted topics, bridging the gap between embedding-based clustering and probabilistic topic models.

Semantic-Driven Topic Modeling

A recent study by Mersha et al. (2024) presents a semantic-driven topic modeling approach that utilizes transformer-based embeddings to capture contextual information. The model generates document embeddings, reduces their dimensions, clusters them based on semantic similarity, and produces coherent topics for each cluster. This method has shown to provide more meaningful topics compared to traditional algorithms.

Applications in Literature Screening

Transformer-based topic modeling algorithms like **BERTopic** have been applied to expedite literature screening processes. For instance, in systematic reviews on peri-implantitis and bone regeneration, BERTopic rapidly identified topic clusters, enabling researchers to filter out unrelated articles efficiently. This application underscores the practical utility of transformer-based models in handling large datasets.

Specialized Transformer Models for Topic Segmentation

1. **Longformer:**

    - Introduced by Beltagy et al. (2020), overcomes the fixed input size limitation inherent in traditional transformer models by employing a sliding-window attention mechanism. In standard transformers, the self-attention layer computes attention scores for every pair of tokens in an input sequence, resulting in quadratic complexity that quickly becomes impractical as sequence lengths increase. Longformer replaces this with a sliding-window approach, which restricts the self-attention computation to a local window around each token. This design change dramatically reduces the computational overhead, scaling the complexity linearly with the sequence length.

      The sliding-window mechanism allows Longformer to efficiently process much longer documents without sacrificing the ability to capture dependencies between nearby tokens. Moreover, Longformer's design includes the option for global attention on a select set of tokens. These tokens often chosen based on their relevance to the task can interact with all other parts of the sequence, ensuring that essential long-range information is not lost. This balance between local and global attention makes Longformer particularly well suited for topic segmentation in long-form content, such as podcast transcripts and books, where maintaining both local coherence and global context is crucial.

      In the context of podcast transcripts, the ability to analyze extended sequences is especially beneficial. Podcasts often contain lengthy, unstructured conversations filled with diverse topics and varying dialogue dynamics. By leveraging Longformer's sliding-window and selective global attention mechanisms, segmentation models can more effectively capture topic boundaries within these extensive transcripts. This results in more precise identification of boundaries without needing to truncate the input or compromise

on context, ultimately enhancing the performance of downstream natural language processing applications like retrieval and summarization.

2. **Hierarchical Transformers:**

   - Hierarchical transformer architectures are designed to process text at multiple granular levels such as sentence-level and paragraph-level representations to capture both local details and global context. Such architectures, as proposed by Gklezakos et al. (2024), first encode smaller text units using a base transformer and then aggregate these representations using a higher-level transformer. This two-stage process allows the model to preserve fine-grained semantic relationships within individual sentences or short spans, while simultaneously maintaining an understanding of the overarching narrative structure across larger text segments. By balancing local semantic cues with global contextual information, hierarchical transformers improve the accuracy in detecting subtle topic boundaries, which is particularly beneficial when segmenting long, unstructured texts such as podcast transcripts.

3. **Multimodal Transformers:**

   - Recent advancements in multimodal transformer architectures have extended the capability of traditional models by integrating information from multiple data sources, most notably text and audio. For instance, Ghinassi et al. (2023) proposed models that leverage pre-trained neural encoders for both modalities to enhance segmentation performance in multimedia data like podcasts. In these systems, the text encoder extracts semantic content from the transcript, while the audio encoder captures prosodic and paralinguistic features such as intonation, rhythm, and pauses that are critical for understanding conversational dynamics. The fusion of these complementary representations enables the model to detect nuanced topic shifts even in noisy environments, where textual information alone may be ambiguous. This multimodal approach thereby enhances the robustness and contextual sensitivity of topic segmentation systems.

Challenges and Limitations

While transformer-based models have revolutionized topic segmentation, they are not without challenges:

1. **Computational Demands:**
    - Transformers require substantial computational resources for training and inference, making them less accessible for researchers with limited resources.
2. **Handling Extremely Long Documents:**
    - Despite advancements like Longformer, transformers still struggle with extremely long sequences, requiring chunking or hierarchical approaches.
3. **Domain Adaptation:**
    - Pretrained transformer models often require fine-tuning to adapt to specific domains, such as podcasts, which can be resource-intensive.

Impact on Topic Segmentation

Transformer-based models have set a new standard for topic segmentation by significantly improving accuracy and robustness. Their ability to handle unstructured, long-form, and conversational text has made them indispensable for modern segmentation tasks. These models have also opened avenues for integrating multimodal data, further enhancing their applicability in real-world scenarios such as podcast analysis and multimedia content segmentation.

The introduction of transformer architectures marked a paradigm shift in natural language processing (NLP), including topic segmentation. Unlike previous models, transformers use self-attention mechanisms to capture dependencies across entire sequences, enabling a deeper understanding of context and relationships between words. These models offer significant advantages in both structured and unstructured text segmentation.

2.2.4 Evaluation of Existing Methods

Xing and Carenini (2021) proposed an innovative approach to unsupervised dialogue topic segmentation using utterance-pair coherence scoring. Their model, grounded in BERT's Next Sentence Prediction (NSP) architecture, demonstrated state-of-the-art results on conversational datasets by predicting

Table 2.1: Comparison of Reproduced Results with Reported Results from Xing and Carenini (2021)

Metric	Model	Reproduced Results	Results in the Paper
P_k Score	TextTiling (TeT)	0.3889	0.40
WindowDiff	TextTiling (TeT)	0.3999	0.40
F1 Score	TextTiling (TeT)	0.6896	0.608
P_k Score	Ours (Full)	0.4218	0.268
WindowDiff	Ours (Full)	0.4718	0.282
F1 Score	Ours (Full)	0.6925	0.776

topic boundaries based on coherence scores between sentence pairs. This approach was particularly effective in multi-speaker dialogue scenarios, where traditional models often struggled.

To validate the reproducibility of their results, a comparative analysis was conducted. Table 2.1 presents a detailed comparison between the originally reported results and the results reproduced during this study.

Analysis and Discussion: As seen in Table 2.1, the reproduced results for the TextTiling model closely align with the reported values, affirming its consistency. This consistency can be attributed to TextTiling's reliance on established lexical cohesion methodologies, which are relatively straightforward to implement and less sensitive to variations in hyperparameters or data preprocessing. Its deterministic and rule-based nature minimizes the potential for discrepancies arising from implementation variability. However, discrepancies in the "Ours (Full)" model indicate potential challenges in replicating the training setup and hyperparameter tuning. These differences underscore the need for careful implementation and suggest avenues for optimization in future work.

2.3 Comparative Analysis of Topic Segmentation Approaches

This section presents a comparative overview of various topic segmentation approaches from the existing literature. To better organize the diversity of input modalities (e.g., dialogue transcripts, text documents, multimodal data) and output tasks (e.g., topic segmentation, shift detection, labeling), these are grouped into four main categories:

- **Dialogue-Based Topic Segmentation**
- **Text-Based Topic Segmentation**

Table 2.2: Dialogue-Based Topic Segmentation Approaches

Index	Research Paper	Input	Output
1	Gao et al. (2023)	Dialogue	Topic segmentation
2	Konigari et al. (2021)	Open-domain Dialogue	Utterance classification
3	Lin et al. (2023b)	Dialogue	Topic shift detection
4	Xie et al. (2021)	Dialogue	Topic shift-aware response generation
5	Lin et al. (2023a)	Dialogue	Topic shift detection

Table 2.3: Text-Based Topic Segmentation Approaches

Index	Research Paper	Input	Output
1	Arnold et al. (2019)	Text	Topic segmentation with labels
2	Yu et al. (2023)	Text	Topic segmentation
3	Lee et al. (2023)	Text	Topic segmentation
4	Çano and Roth (2022)	Text	Topic detection
5	Bai et al. (2023)	WikiSection Text	Topic segmentation

- **Multi-Person and Multimodal Topic Segmentation**
- **Special Task-Specific Topic Segmentation**

The following tables enumerate the most relevant works in each category.

In Table 2.2, the approaches listed are specifically designed for dialogue-based scenarios. These often focus on detecting shifts in topic or speaker initiatives within natural, open-domain conversations, where the structure is often less predictable than in monologue-style texts.

Table 2.3 highlights methods for *single-speaker or monologue* text segmentation. These approaches commonly target academic texts, news articles, or Wikipedia pages, where the structure and vocabulary are more formal, but can still pose challenges for coherent segmentation when documents are lengthy.

Table 2.4 covers works that consider multiple speakers or multiple modalities (e.g., combining text transcripts with audio or visual cues). These studies recognize that topic boundaries in group conversations and multimedia content are heavily influenced by speaker dynamics, prosodic features, and possibly visual signals.

Finally, Table 2.5 focuses on more specialized or hybrid tasks that involve topic segmentation as one component of a broader pipeline (e.g., advanced

Table 2.4: Multi-Person and Multimodal Topic Segmentation

Index	Research Paper	Input	Output
1	Solbiati et al. (2021)	Text(meeting transcripts)	Topic segmentation
2	Ghinassi et al. (2023)	Audio+Text	Topic segmentation
3	Aquilina et al. (2023)	Audi+Text	Enhanced topic segmentation

Table 2.5: Special Task-Specific Topic Segmentation

Index	Research Paper	Input	Output
1	Li et al. (2022)	Transformer attention patterns	Interpretation of attention for summarization and topic segmentation
2	Tannous et al. (2023)	Text(Resumes)	Topic detection
3	Xia and Wang (2023)	Text	Topic segmentation and labeling

summarization, heading detection, or segment labeling). These studies often employ unique architectures or attention-based heuristics that are closely tied to specific domains such as resumes, specialized domains, or interpretability use-cases.

Hence, these four categories illustrate the range of data types and tasks that the research community has addressed, underscoring the importance of choosing or adapting segmentation methods to match the specific constraints and objectives of each domain.

2.4 Evaluation Metrics for Topic Segmentation

Evaluation of topic segmentation methods is a crucial aspect of the research, as it directly informs the effectiveness of the algorithms in practical applications. Traditionally, metrics such as P_k, WindowDiff, and F1 score have been widely adopted:

- **P_k and WindowDiff:** These metrics measure the proportion of incorrectly placed boundaries by comparing a model's segmentation with a reference segmentation (Xing and Carenini, 2021). Although they provide a quantitative basis for comparison, they focus primarily on structural accuracy and may not fully capture the semantic coherence or practical

utility of the segments.

- **F1 Score:** The F1 score, balancing precision and recall, helps evaluate how well a segmentation model identifies true topic boundaries. However, its effectiveness can vary depending on the granularity of the segmentation (Hearst, 1997).
- **Emerging Evaluation Approaches:** Recent studies, such as those by (Yu et al., 2023), advocate for more user-centric evaluation metrics that consider listener satisfaction, navigation efficiency, and the overall usefulness of segmentation in real-world applications. These metrics are crucial for podcast segmentation, where the end-user experience is a critical measure of success.

Recent trends in evaluation emphasise the need for user-centric metrics that account for the practical utility of the segmented output. Future methodologies may incorporate measures of listener satisfaction, navigation efficiency, or task-specific performance (e.g., the effectiveness of chapter-based search functionalities). Such advancements in evaluation are critical for bridging the gap between theoretical performance and real-world applicability.

2.5 Research Gaps and Future Directions in Topic Segmentation

Despite significant progress in topic segmentation—from rule-based methods to advanced transformer-based models—several research gaps remain:

- **Insufficient Contextual Modeling:** Although modern models capture long-range dependencies better than earlier methods, fully understanding the nuances of conversational speech in podcasts still poses challenges.
- **Evaluation Limitations:** Existing evaluation metrics often fail to capture listener-centric aspects, such as the practical utility of chapter segmentation for navigation and content discovery.
- **Data Scarcity and Quality:** The limited availability of high-quality, annotated podcast transcripts constrains the training and evaluation of segmentation models. Robust methods for noise reduction and data normalization are needed.
- **Computational Efficiency:** Processing lengthy podcast transcripts in real time remains computationally demanding, even with efficient trans-

former variants. More scalable solutions are required for widespread practical deployment.

We address these gaps by exploring promising directions for future research. We integrate unsupervised and semi-supervised learning techniques to overcome the challenges posed by limited labeled data. Additionally, we develop novel evaluation frameworks that prioritize end-user experience, providing a more holistic assessment of segmentation quality. Finally, we leverage interdisciplinary insights from audio signal processing and human-computer interaction to inspire new approaches for model design and deployment.

Chapter 3

Methodology

This chapter presents the design and implementation of various approaches for topic segmentation of podcast transcripts, including classical unsupervised methods and transformer-based models. The objective is to transform unstructured and lengthy podcast transcripts into semantically coherent segments, thereby enhancing content navigation, summarization, and downstream processing.

To achieve this, we evaluate and compare multiple methods, including traditional segmentation algorithms and large language models (LLMs), assessing their effectiveness in segmenting podcast transcripts. The methodology follows a structured process comprising four major stages: (1) Creating a Dataset for Topic Segmentation in Podcasts, (2) Manual Annotation, (3) Automated Topic Segmentation.

Detailed descriptions of each stage are provided in the following sections.

The primary goal of this research is to decompose podcast transcripts into meaningful topics, facilitating efficient information retrieval and improved comprehension. Podcast data is inherently unstructured and conversational, often containing informal language, digressions, and non-linear narrative elements that challenge conventional segmentation techniques. Traditional methods, which largely rely on fixed rules or simplistic lexical cohesion measures, frequently fail to capture the nuanced semantic transitions present in such data.

To address these challenges, the proposed method integrates advanced LLM-based techniques with classical segmentation approaches. The LLM component leverages deep semantic representations to identify topic shifts and extract underlying themes, while classical methods (e.g., TextTiling) contribute computational efficiency and interpretability. This hybrid approach ensures robustness and adaptability across diverse podcast genres and audio qualities. The chapter details the comprehensive process employed for data preparation,

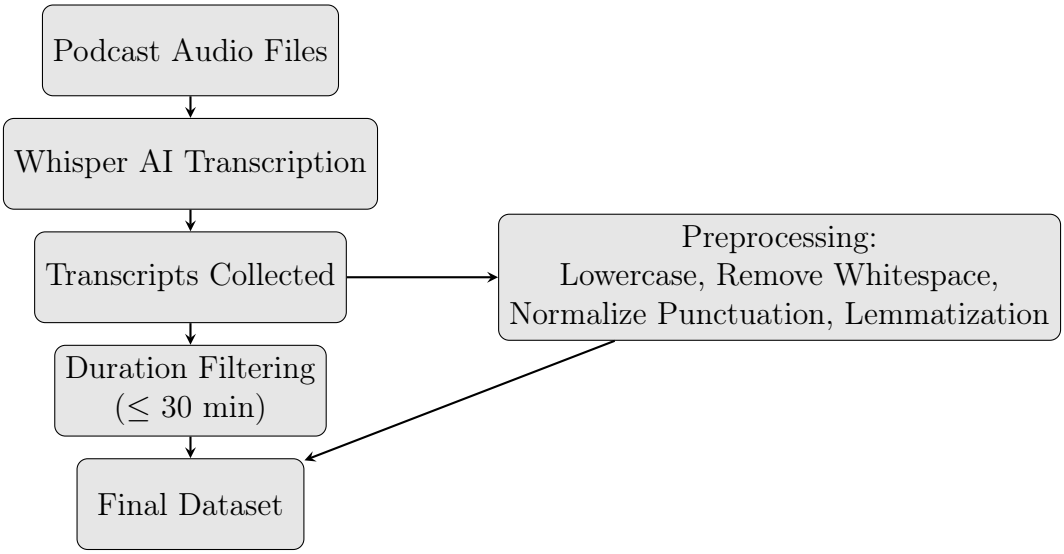

Figure 3.1: Data Collection and Preprocessing Workflow. Raw podcast audio is transcribed via Whisper AI, then transcripts are preprocessed and filtered (by duration) to form the final dataset.

manual annotation, the development of automated segmentation methods, and the evaluation of the system's performance.

3.1 Creating a Dataset for Topic Segmentation in Podcasts

3.1.1 Data Collection

Figure 3.1 presents the comprehensive data collection and preprocessing pipeline. We prepare raw podcast audio files are first transcribed using the OpenAI Whisper model, which has been demonstrated to achieve state-of-the-art transcription accuracy and robustness in challenging real-world environments (Radford et al., 2022). The Whisper model employs a transformer-based encoder-decoder architecture, trained on 680,000 hours of multilingual and multitask supervised data, which enables it to handle a wide range of accents, background noise, and language nuances effectively.

For this study, 100 podcast audio files were processed through the Whisper model to generate corresponding text transcripts. These transcripts preserve the inherent characteristics of spoken language—including informal expressions and natural disfluencies—and provide the raw input for subsequent segmentation. To ensure that the manual annotation workload remained manageable

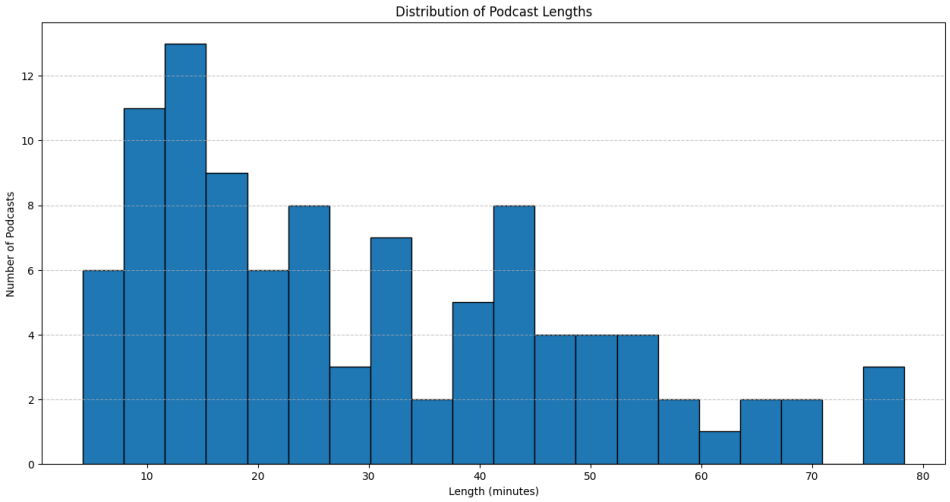

Figure 3.2: Histogram of podcast durations (in minutes) for the collected dataset.

and that only the most relevant data were processed, the dataset was filtered to include only those podcasts with durations of 30 minutes or less. An analysis of the podcast length distribution confirmed that a significant subset met this duration criterion, as illustrated by the histogram in Figure 3.2.

In the workflow diagram (Figure 3.1), the entire process is depicted as a series of steps: raw audio files are first transcribed using Whisper AI, then the resulting transcripts are preprocessed (including conversion to lowercase, removal of extraneous whitespace, punctuation normalization, and lemmatization) before being filtered by duration. This dual-path process (direct collection and simultaneous preprocessing) culminates in the final curated dataset, which retains the semantic richness of the original audio while being optimized for manual annotation and automated segmentation.

3.1.2 Preprocessing

A minimal yet effective preprocessing strategy was adopted to preserve the integrity and natural structure of the spoken content. Focusing on shorter podcasts allowed a balance to be achieved between maintaining topic diversity and keeping the manual annotation process manageable, thereby supporting the development of a robust, LLM-based topic segmentation framework.

Minimally invasive preprocessing was considered essential in order to retain the full context inherent in podcast audio. Consequently, the raw transcripts—complete with introductions, advertisements, and outros—were pre-

served without significant alterations. It was also observed that the transcripts contain numerous filler expressions, such as "mm-hmm", which occur continuously throughout the conversations. Although these disfluencies introduce a level of noise, they are retained in order to preserve the natural cadence and authenticity of the spoken discourse. Such filler words may provide subtle cues about topic transitions and speaker engagement, and thus are considered valuable for downstream segmentation tasks.

The standard preprocessing steps applied to the transcripts included:

- **Converting text to lowercase:** Ensures that comparisons and subsequent processing are case-insensitive, which is crucial for consistent tokenization and lexicon matching.
- **Removing extraneous whitespace and normalizing punctuation:** Eliminates inconsistencies and formatting errors that can arise during transcription, thereby improving the performance of downstream tokenization and segmentation algorithms.
- **Applying lemmatization:** Reduces words to their base or dictionary form, minimizing vocabulary redundancy. This step is particularly beneficial for semantic analysis and for constructing effective representations for LLM-based segmentation models.

This preprocessing approach not only preserves the natural cadence and structural markers of the conversation but also formats the data in a manner suitable for both manual annotation and automated segmentation. As a result, the data remains faithful to the original audio while being streamlined enough for effective LLM processing and topic segmentation. Figure 3.3 illustrates the preprocessing pipeline, visually outlining the transformation of raw transcripts into preprocessed text.

3.1.3 Challenges in Data Collection

Data collection for podcast segmentation presents several significant challenges. Two primary challenges were identified during this phase:

- **Length Variability:** The initial dataset comprised podcasts with durations ranging from 10 to over 100 minutes. Such variability poses challenges for both manual annotation and automated segmentation, as longer podcasts demand more extensive effort to accurately segment and may contain multiple, interleaved topics. To ensure consistency and manageability, a duration threshold was imposed to restrict the dataset to podcasts lasting between 10 and 30 minutes. This decision, supported by

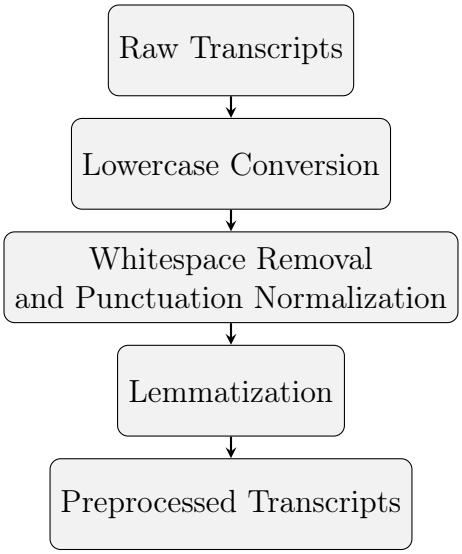

Figure 3.3: Preprocessing Pipeline: Raw transcripts are converted to lowercase, cleaned of extraneous whitespace and normalized for punctuation, then lemmatized to produce preprocessed transcripts.

an analysis of the podcast duration distribution (see Figure 3.2), is consistent with approaches used in previous research (Hearst, 1997; Clifton et al., 2020).

- **Selection Criteria:** Balancing topic diversity with annotation feasibility necessitated a careful filtering process based on both duration and content relevance. Metadata such as episode descriptions and contextual cues were employed to ensure that the selected podcasts provided rich, coherent topical content. This filtering strategy helped to exclude episodes with low-quality or off-topic content, thereby improving the representativeness and reliability of the dataset. Similar challenges and filtering approaches have been reported in large-scale podcast corpus studies (Clifton et al., 2020).

Additional challenges encountered during data collection included the presence of noisy transcription outputs, speaker variability, and inconsistent metadata. These factors further complicated the selection process and required robust preprocessing and filtering techniques to ensure that the final dataset remains both manageable and representative of the diverse nature of podcast audio.

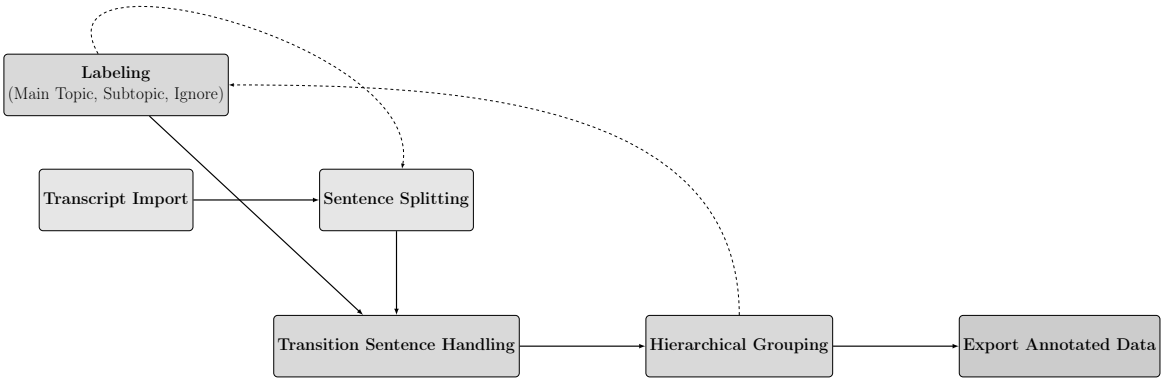

Figure 3.4: Manual annotation workflow for topic segmentation. Transcripts are imported and split into sentences. Labeling is applied to categorize content into main topics, subtopics, and ignore tags, followed by transition sentence handling and hierarchical structuring. Finally, the annotated data is exported. Dashed arrows represent feedback loops for iterative refinement.

3.2 Manual Annotation of Topic Segments in Podcasts

Manual annotation plays a critical role in establishing the ground truth necessary for evaluating automated segmentation algorithms. A single annotator performed the labeling of topic boundaries within the podcast transcripts using a hierarchical labeling scheme. This rigorous process ensures that the annotated data accurately reflects the natural structure of the discourse, thereby providing a robust benchmark for assessing the performance of segmentation methods.

The manually annotated transcripts serve as the gold standard against which automated systems are compared. They guide the development and refinement of computational models and help uncover intrinsic challenges such as ambiguities in topic transitions or overlapping themes that are common in natural spoken language.

A workflow diagram outlining the manual annotation process is presented in Figure 3.4. This diagram illustrates the sequential steps: transcripts are first imported, then segmented into candidate spans, followed by manual labeling according to the annotation guidelines, and finally organized into a hierarchical structure to form the annotated gold standard.

Figure 3.5 provides a visual example of how the hierarchical labeling scheme is applied to a typical podcast transcript. In this illustration, the text is divided

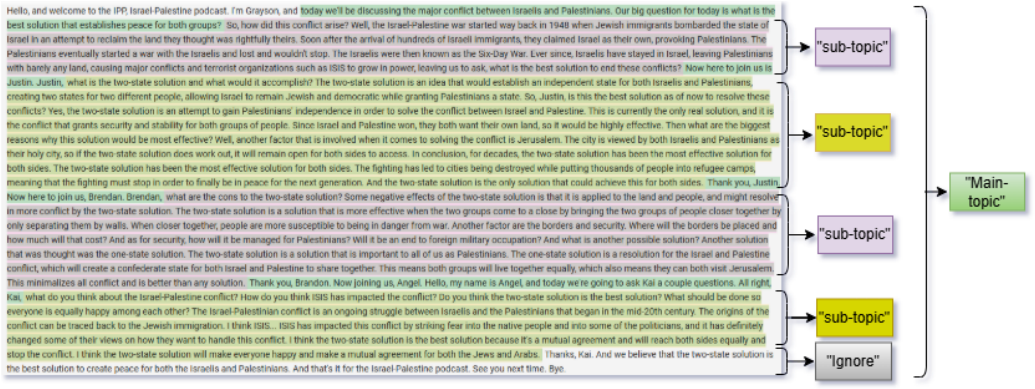

Figure 3.5: Example of a manually annotated podcast transcript. Main topics, subtopics, and ignore segments are color-coded to reflect their respective roles in the conversation flow.

into color-coded segments to distinguish **Main Topic**, **Subtopic**, and **Ignore** sections.

- **Main Topic:** Represents the primary theme of discussion within a segment. These are the key ideas or central points around which the conversation is structured. - **Subtopic:** Denotes a related but secondary discussion point that provides further elaboration, supporting details, or contextual extensions of the main topic. - **Ignore:** Refers to content that does not contribute to topic segmentation, such as filler words, incomplete thoughts, off-topic remarks, or conversational digressions.

Sentences that serve as bridges between topics are included in both adjacent segments, capturing the often gradual and overlapping nature of spoken discourse. Combining a clear labeling scheme with a structured annotation process makes the ground truth data a reliable reference for evaluating the accuracy of automated segmentation methods.

3.2.1 Annotation Guidelines

To maintain high-quality and consistent annotations, detailed guidelines were developed to direct the annotator in capturing both the overarching narrative structure and the finer nuances of content shifts. The main components of the guidelines are as follows:

- **Main Topic:**
  - Each podcast is assigned an overarching theme that encapsulates the primary discussion. For example, a podcast exploring financial

markets may have the main topic labeled as "Economic Trends in Europe."

  - This label serves as the primary anchor point for segmentation, reflecting the core context of the conversation.

- **Subtopic:**

  - Within each main topic, sections that explore specific aspects or arguments are annotated as subtopics (e.g., "Inflation," "Tax Policies," or "Monetary Policy").
  - Subtopics capture granular shifts in focus, enabling a multi-level evaluation of segmentation performance at both the coarse (main topic) and fine (subtopic) levels.

- **Ignore:**

  - Non-content elements such as introductions, advertisements, interludes, and any peripheral content that do not contribute to the core discussion are marked with the label `Ignore`.
  - Excluding these sections ensures that segmentation metrics focus solely on the substantive parts of the transcript.

- **Transition Sentences:**

  - Sentences that function as bridges between distinct topics are annotated as belonging to both adjacent segments.
  - This dual-labeling captures the inherent ambiguity in natural discourse, where topic shifts are often gradual rather than abrupt.

- **Hierarchical Grouping:**

  - The annotation framework organizes subtopics hierarchically under their corresponding main topics, mirroring the natural discourse structure.
  - This grouping facilitates multi-level evaluation of segmentation and supports detailed analysis of topic boundaries.

So, this manual annotation protocol lays a strong foundation for evaluating automated segmentation methods by ensuring that the ground truth data is both reliable and representative of the nuanced structure of spoken discourse.

3.2.2 Annotation Execution

The manual annotation process was performed collaboratively by two annotators—the author and the supervisor—to generate a robust ground truth dataset for evaluating automated segmentation. The process was executed according to the following detailed steps:

1. **Transcript Segmentation:** Each transcript was divided into discrete text spans, each consisting of 1 to 5 sentences. This granularity was chosen to capture individual coherent ideas, ensuring that each span represents a single claim, argument, or explanatory unit.

2. **Label Assignment:** Each text span was then assigned one of three categorical labels based on its content:

   - **Main Topic:** Denotes the overarching theme of the podcast (e.g., "Economic Trends in Europe").
   - **Subtopic:** Represents a more focused discussion or an elaboration on a facet of the main topic (e.g., "Inflation Dynamics" or "Tax Policy Reforms").
   - **Ignore:** Applied to non-core content such as advertisements, introductions, and other peripheral elements.

3. **Transition Sentence Handling:** Sentences that serve as transitions between two distinct topics were included in both adjacent segments. This overlapping ensures contextual continuity and a smoother flow between topics.

4. **Descriptive Labelling:** In addition to the categorical labels, each segment was assigned a concise descriptive label (e.g., "Economic Impact of Immigration") that summarizes the core discussion of that segment. These labels are critical for subsequent analysis and interpretation.

5. **Collaborative Quality Assurance:** Throughout the annotation process, the annotators engaged in regular discussions to resolve ambiguities and ensure consistent application of the guidelines across the entire dataset. This collaborative approach improved the reliability of the annotations and minimized subjectivity.

This structured annotation execution procedure not only ensures consistency and clarity in the manual segmentation process but also provides a reliable basis for the evaluation of automated topic segmentation algorithms which will be discussed in the upcoming sections.

3.3 Automated Topic Segmentation Methods

To scale beyond manual annotation, three automated segmentation approaches were implemented: TextTiling (Baseline), LLM-Based Topic Extraction with Similarity Thresholding, and a Transformer-Based Model with BIO Labeling. Each method is tailored to address the challenges of segmenting podcast transcripts, which are characterized by informal language, disfluencies, and varied topical shifts.

3.3.1 TextTiling (Baseline)

TextTiling is a classical segmentation algorithm that relies on lexical cohesion to detect shifts in topics by analyzing patterns in word distributions. In this work, the TextTiling method was adapted to address the unique challenges posed by podcast transcripts, which are often noisy and exhibit informal language patterns. The following detailed workflow was implemented to perform topic segmentation using TextTiling:

Implementation Workflow

1. **Preprocessing:** Artificial paragraph breaks were inserted after every five sentences to simulate a structured topic flow. This artificial segmentation allows for treating continuous stretches of text as cohesive blocks. Prior to this, transcripts were converted to lowercase and lemmatized to ensure uniform token representation and reduce vocabulary redundancy.

2. **Tokenization:** The preprocessed transcripts were segmented into individual sentences using a robust sentence tokenizer. This step preserves natural sentence boundaries, ensuring that the algorithm works with linguistically meaningful units.

3. **Sliding Window Comparison:** A fixed sliding window covering five consecutive sentences was applied over the tokenized transcript. For each window, word frequency distributions were computed and cosine similarity was calculated between adjacent windows. A significant drop in cosine similarity between two neighboring windows was interpreted as an indicator of a topic boundary, based on the assumption that semantically coherent text blocks will exhibit high lexical similarity.

4. **Output Generation:** The boundaries detected through the sliding window comparison were recorded in a structured format, allowing for systematic evaluation against the manually annotated gold standard. This

structured output facilitates direct comparison with the ground truth during the evaluation phase.

Strengths and Limitations

Strengths:

- **Language-Agnostic and Efficient:** The method relies solely on lexical statistics, making it inherently language-agnostic and computationally efficient. This efficiency enables it to scale to large podcast datasets without requiring extensive computational resources.
- **Unsupervised Approach:** As an unsupervised technique, TextTiling does not require any labeled training data. This makes it particularly attractive for domains where annotated data is scarce or expensive to obtain, and it can be readily applied to diverse datasets without domain-specific tuning.

Limitations:

- **Sensitivity to Noise:** The method is highly sensitive to noise present in spoken transcripts, such as frequent filler words (e.g., "mm-hmm") and other disfluencies that naturally occur in conversational speech. These elements can distort lexical similarity computations and lead to inaccurate boundary detection.
- **Fixed Window Dependency:** TextTiling's reliance on a fixed sliding window (e.g., five sentences) can result in over-segmentation or under-segmentation when topic transitions are not uniformly distributed across the transcript. This rigid structure may not adequately capture the fluid nature of topic shifts in podcast data.
- **Limited Semantic Depth:** Since the approach is based on surface-level lexical cohesion, it has a limited capacity to capture deeper semantic relationships between sentences. This limitation is particularly critical in podcast transcripts, where nuanced and context-dependent topic transitions often occur.

Hence, while TextTiling offers a fast and language-agnostic solution that is easy to deploy across large datasets, its sensitivity to noise and reliance on fixed window sizes may reduce its effectiveness in capturing the rich, fluid semantic structures of podcast transcripts. These limitations have motivated the exploration of more context-aware methods, such as LLM-based topic extraction and transformer-based BIO labeling, which aim to overcome these challenges by leveraging deep semantic representations.

3.3.2 LLM-Based Topic Extraction with Similarity Thresholding

This approach leverages large language models to extract key topics from each podcast transcript and subsequently assigns sentences to these topics based on their semantic similarity. In this work, a variant of the Llama model (model: `llama3-8b-8192`[1]) is employed for generating concise and distinct topic labels, while the `all-mpnet-base-v2` model[2] is used to produce high-quality semantic embeddings for both the extracted topics and individual sentences. This combination enables the capture of deep contextual nuances and ensures robust topic segmentation even in the presence of informal and noisy spoken language.

Implementation Workflow

The complete process comprises the following steps:

1. **Topic Extraction:** The full transcript is input into a large language model, which is prompted to extract up to five primary topics that best summarize the overall content. The prompt is carefully crafted to instruct the model to return a list of broad and distinct topic names. This step exploits the LLM's capability to understand context and generate coherent thematic labels.

2. **Sentence Segmentation:** The transcript is then divided into individual sentences to ensure that natural sentence boundaries are preserved. Accurate sentence segmentation is essential for aligning each sentence with its corresponding topic.

3. **Embedding Generation:** Semantic embeddings for both the extracted topics and the individual sentences are computed using the `all-mpnet-base-v2` model. These embeddings capture the deep semantic properties of the text, facilitating robust similarity computations.

4. **Assignment via Similarity Thresholding:** For each sentence, cosine similarity is computed between its embedding and the embeddings of the extracted topics. A sentence is assigned to a particular topic if its similarity score exceeds a predefined threshold. Sentences that do not meet this threshold for any topic are marked as miscellaneous.

[1]The variant of the Llama model (llama3-8b-8192) used for topic extraction is based on Meta's Llama architecture. For more details, see `https://ai.facebook.com/blog/large-language-models-llama/`.

[2]The `all-mpnet-base-v2` model is available on Hugging Face at `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`.

5. **Refinement:** To improve segmentation accuracy, an iterative refinement process is applied. In each iteration, topic prototypes are recalculated by averaging the embeddings of all sentences currently assigned to a topic. The similarity-based assignment is then repeated for a fixed number of iterations, ensuring that the topic boundaries are accurately captured and that the assignment of sentences converges.

Discussion

This LLM-based approach is particularly suited for the segmentation of podcast transcripts due to its ability to manage the informal and often noisy nature of spoken content. The use of the Llama model for topic extraction leverages advanced natural language understanding to derive meaningful thematic labels, while the `all-mpnet-base-v2` model provides rich semantic embeddings that effectively capture contextual nuances. This integration enables reliable similarity computations, even in the presence of filler words and disfluencies, and supports iterative refinement to optimize topic assignments.

Figure 3.6 illustrates the complete workflow for LLM-based topic extraction with similarity thresholding. The diagram visually represents the sequence of processing steps:

- It begins with the *Transcript Input*, which is processed to extract topics using a large language model.
- The transcript is then segmented into sentences, preserving natural boundaries.
- Next, semantic embeddings for both sentences and extracted topics are generated.
- The system computes cosine similarity between the sentence embeddings and topic embeddings, and assigns sentences to topics if the similarity exceeds the threshold.
- Finally, an iterative refinement loop is implemented, where the topic prototypes are updated based on current assignments to enhance the segmentation accuracy.

This diagram provides a clear and concise overview of the segmentation pipeline, highlighting both the sequential processing steps and the feedback loop essential for iterative refinement.

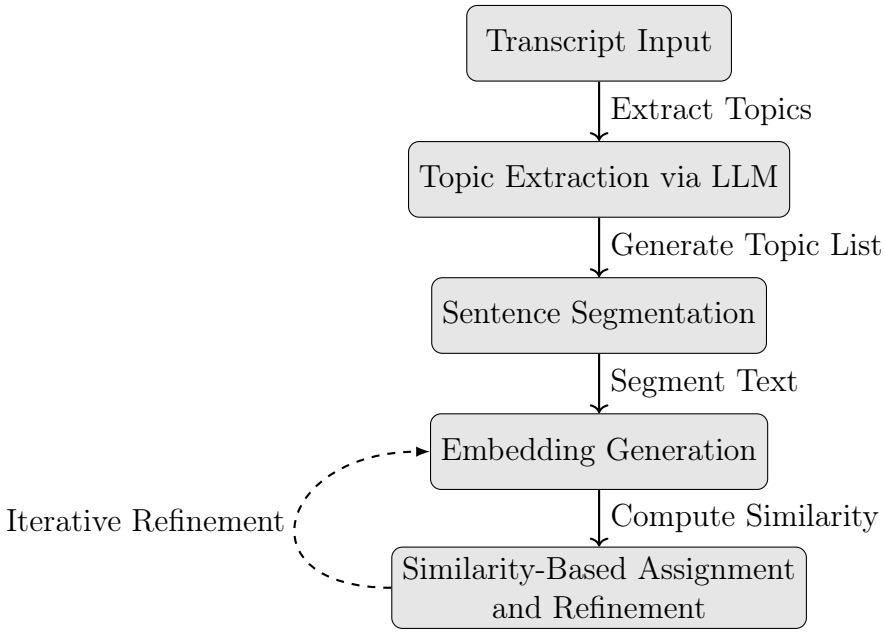

Figure 3.6: Workflow for LLM-Based Topic Extraction with Similarity Thresholding. The process begins with the transcript input, followed by topic extraction using an LLM, sentence segmentation, embedding generation, and finally similarity-based assignment with iterative refinement to optimize topic boundaries.

3.3.3 Transformer-Based Model with BIO Labeling

A supervised segmentation approach was also employed using a transformer-based model that is fine-tuned for sequence labeling on manually annotated podcast transcripts. This method utilizes a DistilBERT-based encoder, combined with a Conditional Random Field (CRF) layer, to predict BIO (Begin, Inside, Outside) tags for each sentence. The BIO tagging scheme is a widely used method in sequence labeling tasks that helps in identifying the boundaries of meaningful segments. Specifically, the tag **B** is used to mark the first sentence of a new topic segment, **I** indicates that the sentence continues the topic, and **O** denotes sentences that do not belong to any topic segment. For example, if a new topic starts with "Economic Trends in Europe," the first sentence discussing this theme would be tagged as **B**; subsequent sentences elaborating on the same topic would be tagged as **I** until a transition is observed, at which point a new **B** tag is applied.

Implementation Workflow

The complete workflow for the transformer-based BIO labeling model comprises the following steps:

1. **BIO Tagging:** Each sentence in the transcript is manually annotated using the BIO scheme. These manually assigned BIO tags serve as the gold standard for supervised training.

2. **Chunking:** To accommodate the input length limitations of transformer models, transcripts are segmented into chunks containing 20 sentences each. This step ensures that each input remains within the maximum token length supported by the model while preserving the sequential structure of the discourse.

3. **Model Training:** A DistilBERT-based model is fine-tuned for sequence labeling using the prepared BIO-annotated data. A CRF layer is added on top of the transformer to capture dependencies between adjacent labels, ensuring that the predicted label sequence adheres to the BIO format. Training is performed using Leave-One-Out Cross-Validation (LOOCV).

4. **Evaluation:** The performance of the model is assessed using standard segmentation metrics such as the F1 score, P_k, and WindowDiff (WD). Predicted BIO sequences are converted into topic boundaries and compared with the gold standard annotations.

Detailed Implementation and Discussion

The training pipeline begins by loading the manually annotated transcripts, where each sentence is labeled according to the BIO scheme. Transcripts are split into sentences using a sentence tokenizer, and then grouped into chunks of 20 sentences to meet the transformer's input requirements. Each sentence is tokenized and processed by DistilBERT[3] to generate contextual embeddings. Mean pooling is applied to obtain sentence-level representations, and positional embeddings are added to preserve the sequential order of sentences within each chunk.

Figure 3.7 illustrates the end-to-end pipeline of the Transformer-Based BIO Labeling approach. It depicts the sequence of steps from loading CSV data, parsing annotations, chunking text, tokenizing with DistilBERT, and incorporating positional embeddings, to training with a CRF layer in a Leave-One-Out Cross-Validation (LOOCV) setup. The iterative LOOCV process, represented by the dashed arrow in the diagram, ensures robust evaluation across multiple data folds.

These sentence embeddings are fed through a linear classifier to produce emission scores for each sentence, which are then decoded by the CRF layer to generate the final BIO label sequence. The CRF layer ensures that the output sequence respects the BIO constraints (for example, an **I** label must follow a **B** or another **I** label). The model is fine-tuned on the gold standard data using LOOCV, allowing each transcript to serve as a validation set in turn. Evaluation is carried out by converting the predicted BIO sequences into topic boundaries and comparing these with the manual annotations using F1, P_k, and WD metrics.

This transformer-based approach, with its ability to leverage deep contextual representations and model sequential dependencies through the CRF layer, is well-suited to capturing the complex, noisy, and informal structure of podcast transcripts. While chunking may limit some long-range context, the overall pipeline effectively identifies topic boundaries and provides reliable segmentation performance.

[3]The pretrained DistilBERT model used in this work is available on Hugging Face at `https://huggingface.co/distilbert-base-uncased`.

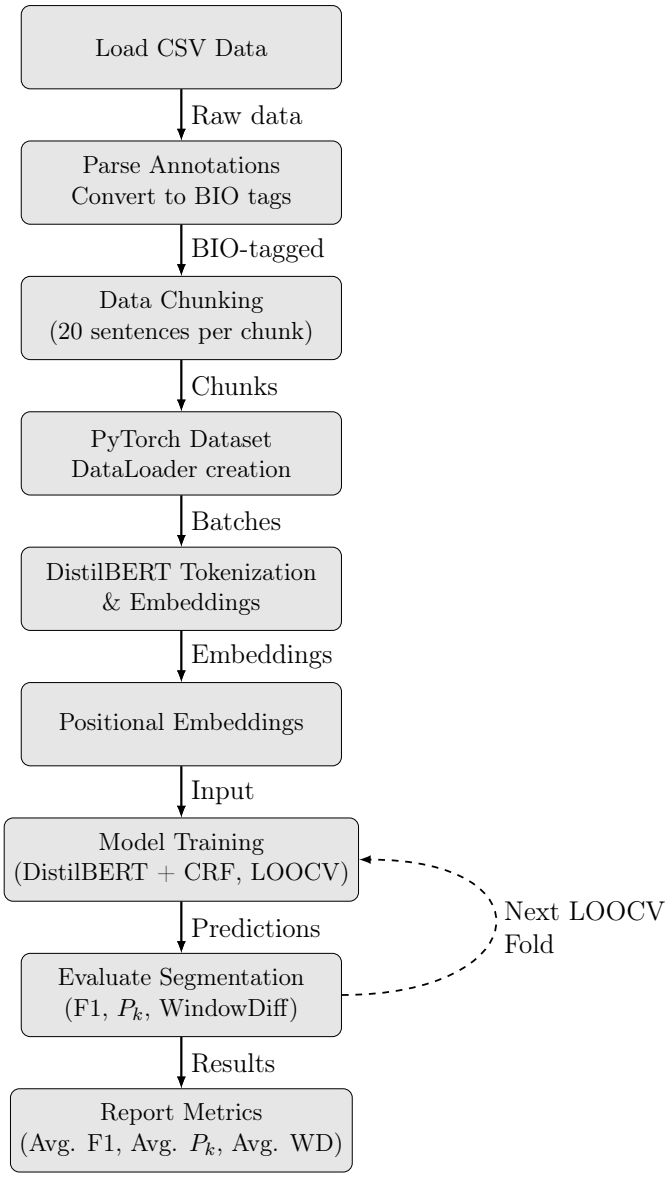

Figure 3.7: Transformer-Based BIO Labeling workflow for podcast transcript segmentation. Data is loaded, annotated, chunked, and tokenized with DistilBERT embeddings and positional encodings. The DistilBERT model with a CRF layer is trained in a LOOCV setting. Model predictions are evaluated using standard metrics (F1, P_k, WindowDiff). The dashed arrow illustrates the iterative LOOCV process, repeating training and evaluation for each data fold.

Summary of Automated Methods

The automated segmentation framework developed in this thesis comprises three distinct approaches, each designed to address specific challenges associated with podcast transcripts:

- **TextTiling:** A classical, unsupervised method that utilizes lexical cohesion and fixed sliding windows to detect topic boundaries. It operates by identifying significant drops in cosine similarity between adjacent text blocks, thus signaling shifts in topic. This method is computationally efficient and language-agnostic, though it is sensitive to noise and rigid window sizes.
- **LLM-Based Topic Extraction:** A deep learning approach that leverages large language models to extract a set of key topics from the full transcript. Semantic embeddings for both the extracted topics and individual sentences are generated using the `all-mpnet-base-v2` model, and sentences are assigned to topics based on a predefined cosine similarity threshold. This method effectively captures contextual nuances and is robust against informal language and filler words.
- **Transformer-Based BIO Labeling:** A supervised segmentation technique that employs a DistilBERT-based model fine-tuned for sequence labeling with BIO (Begin, Inside, Outside) tags. By modeling the sequential dependencies of sentence labels using a Conditional Random Field (CRF) layer, this approach precisely identifies topic boundaries. Chunking of transcripts into manageable segments further facilitates processing while preserving local context.

Each method was specifically designed to overcome the challenges inherent in processing podcast transcripts, such as the presence of disfluencies, variable topical structures, and noise from automatic speech recognition. The strengths and limitations of these approaches are further analyzed through rigorous evaluation using metrics such as the F1 score, P_k, and WindowDiff, which are discussed in detail in a subsequent section.

Chapter 4

Evaluation and Results

Evaluation is a crucial component in validating the effectiveness and robustness of topic segmentation methods, particularly for podcasts, which exhibit diverse linguistic and thematic complexity. This chapter provides a comprehensive analysis and interpretation of the results obtained from the automated segmentation methods implemented in this thesis—TextTiling, LLM-based Similarity Thresholding, and Transformer-based BIO labeling. The primary objective of this evaluation is to systematically compare each method's ability to accurately identify coherent topic boundaries within podcast transcripts, which are inherently unstructured, conversational, and noisy.

The evaluation is performed on the manually annotated dataset, which serves as the gold standard, enabling a precise assessment of each segmentation approach. Given the substantial complexity and variability in the spoken podcast content, evaluation metrics have been selected to provide both quantitative measures and qualitative insights. The quantitative evaluation relies on well-established segmentation metrics, including the F1 Score, P extsubscriptk, and WindowDiff (WD). These metrics are chosen because they provide complementary insights into the methods' performance, capturing both the accuracy of boundary predictions and robustness to variations in segment lengths. Additionally, a qualitative error analysis is conducted to explore and explain the types and sources of segmentation errors encountered.

To facilitate a clear understanding of the evaluation procedure and results, this chapter is structured systematically as follows: First, the evaluation metrics are thoroughly defined and their relevance to topic segmentation is explained. Next, the experimental setup, including dataset details, evaluation protocol (e.g., Leave-One-Out Cross-Validation), and specific hyperparameters used in each method, are described. Subsequently, the detailed quantitative results obtained from each method are presented, supported by relevant visualizations such as bar plots and charts. An in-depth qualitative error analysis

follows, highlighting frequent segmentation errors and exploring their potential causes and implications. Finally, the chapter concludes with a comprehensive discussion summarizing the strengths, limitations, and practical implications of each segmentation approach, along with recommendations for future research directions.

This structured evaluation ensures not only a rigorous assessment of the implemented segmentation techniques but also provides valuable insights into their practical applicability for enhancing content navigation, summarization, and downstream NLP tasks in the context of podcast transcript analysis.

4.1 Experimental Setup

To ensure a systematic and thorough evaluation of the implemented automated segmentation methods, a detailed and structured experimental setup was established. This setup includes clearly defined datasets, hardware and software configurations, model hyperparameters, and the evaluation protocol.

4.1.1 Dataset

The dataset utilized in this research consists of podcast transcripts that were collected and processed through the Whisper AI model, yielding highly accurate textual representations suitable for segmentation tasks. Initially, a dataset comprising 100 podcast transcripts was generated from podcast audio files. However, to facilitate manageable and accurate manual annotations and subsequent automated processing, a subset of 30 podcast transcripts, each between 10 to 30 minutes in length, was selected. This filtering criterion was informed by an analysis of podcast length distribution, which indicated optimal balance between annotation feasibility and topic diversity.

4.1.2 Evaluation Metrics

The performance of the segmentation methods was assessed using three established metrics: F1 Score, P_k, and WindowDiff (WD). These metrics collectively provide insights into the segmentation quality from different perspectives:

F1 Score: The F1 score measures the harmonic mean between precision (the fraction of correctly identified boundaries among all boundaries identified by the model) and recall (the fraction of correctly identified boundaries among all actual boundaries). It is particularly valuable in segmentation tasks as it balances boundary detection accuracy and completeness, making it a comprehensive metric for evaluating topic segmentation performance.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.1}$$

Pk Metric: P_k is specifically designed to measure segmentation errors, defined as the probability that two points randomly selected from the text at a given window distance k are incorrectly identified as either belonging to the same segment or to different segments. A lower P_k score indicates better performance, demonstrating fewer segmentation errors. This metric is widely used because it directly addresses the alignment quality of predicted and actual segmentations.

The P_k metric is computed as:

$$Pk = \frac{1}{N-k} \sum_{i=1}^{N-k} \mathbb{I}(R(i, i+k) \neq P(i, i+k)) \tag{4.2}$$

where:

- N is the total number of text units (e.g., sentences or tokens),
- k is the window size, typically set to half the average reference segment length,
- $R(i, i+k)$ indicates whether two points in the reference segmentation belong to the same segment,
- $P(i, i+k)$ indicates whether the predicted segmentation assigns them to the same segment,
- $\mathbb{I}$ is the indicator function that returns 1 if the prediction is incorrect and 0 otherwise.

WindowDiff Metric: WindowDiff (WD) is another widely adopted metric for evaluating segmentation accuracy, improving upon P_k by considering the number of boundaries in a sliding window of fixed length. Similar to P_k, a lower WindowDiff score signifies better segmentation. WindowDiff captures boundary placement more precisely, making it a complementary and reliable measure alongside the F1 score and P_k.

The WindowDiff metric is defined as:

$$WD = \frac{1}{N-k} \sum_{i=1}^{N-k} \left| B_R(i, i+k) - B_P(i, i+k) \right| \tag{4.3}$$

where:

- $B_R(i, i+k)$ is the number of boundaries in the reference segmentation within the window,
- $B_P(i, i+k)$ is the number of boundaries in the predicted segmentation within the window.

Both P_k and WindowDiff offer valuable insights into segmentation performance, with WindowDiff being particularly advantageous in cases where exact boundary alignment is essential. These metrics serve as essential evaluation tools for topic segmentation models, helping to ensure consistency and reliability in boundary detection.

4.2 Results and Comparative Analysis

In this section, detailed results and analyses are presented for each of the implemented segmentation methods—TextTiling, LLM-based Similarity Thresholding, and Transformer-based BIO Labeling. The evaluation metrics (F1 Score, P_k, and WindowDiff) were computed across all annotated podcast transcripts, providing a thorough comparison of their performance.

4.2.1 Quantitative Results Analysis

Table 4.1 summarizes the average performance metrics obtained from the evaluation of the segmentation methods on the selected podcast transcripts.

Method	**F1 Score**	**P_k**	**WindowDiff**
TextTiling	0.53	0.44	0.45
LLM Similarity Thresholding	0.72	0.29	0.31
Transformer BIO Labeling	0.47	0.57	0.68

Table 4.1: Quantitative Evaluation Results of Automated Segmentation Methods.

In addition to the quantitative summary provided in Table 4.1, a detailed examination of each segmentation method's strengths and limitations reveals how their underlying algorithms perform when tasked with identifying topic boundaries in noisy, conversational podcast transcripts. TextTiling, serving as a traditional baseline, is compared against both the LLM Similarity Thresholding and Transformer-based BIO Labeling methods to highlight key differences.

TextTiling (Baseline): TextTiling exhibits moderate segmentation quality, as reflected by an F1 Score of 0.53, a P_k score of 0.44, and a WindowDiff of 0.45.

The method benefits from its reliance on lexical cohesion: abrupt changes in word usage typically lead to boundary detection, enabling it to capture overt topic shifts. However, subtle or overlapping topics pose difficulties. When adjacent segments share common vocabulary or transition gradually, TextTiling can mistakenly merge them, resulting in a higher P_k and WindowDiff. These errors become more pronounced in podcasts where informal language, fillers, and frequent digressions blur clear lexical cues.

LLM Similarity Thresholding: By contrast, the LLM-based Similarity Thresholding method substantially outperforms TextTiling on all reported metrics (F1 Score of 0.72, P_k of 0.29, WindowDiff of 0.31). Its reliance on contextual embeddings rather than raw lexical overlap allows it to better handle semantically subtle boundaries, such as transitions involving synonyms or related concepts. Nevertheless, careful calibration of the similarity threshold remains crucial. A low threshold risks over-segmentation (excessive topic assignments), while a high threshold can yield under-segmentation by discarding sentences with moderate similarity. Despite these trade-offs, the method's semantic sensitivity proves especially advantageous in podcasts containing nuanced or overlapping discussions.

Transformer-Based BIO Labeling: The Transformer-based BIO Labeling approach, which achieved an F1 Score of 0.47, a P_k of 0.57, and a WindowDiff of 0.68, performs variably in comparison to TextTiling. Although it leverages powerful contextual representations via DistilBERT and a CRF layer, its success hinges on the availability of high-quality annotated data. The supervised nature of the method can be a strength as once the model is well-trained, it captures complex transitions but any inconsistency in the manual labels or insufficient coverage of training examples can degrade performance. Furthermore, the hierarchical or multi-speaker structures often found in podcasts may require more specialized labeling schemes to fully capitalize on the model's potential.

Overall Comparison: Figure 4.1 visually depicts the relative performance of these three methods. While TextTiling provides a computationally lightweight baseline, it struggles with nuanced topic changes. The LLM Similarity Thresholding approach addresses many of these shortcomings by incorporating deeper semantic understanding, leading to the most balanced performance across F1, P_k, and WindowDiff. The Transformer-based BIO Labeling method offers strong theoretical advantages through contextual embeddings and supervised learning but can be hindered by the complexity and variability of podcast data.

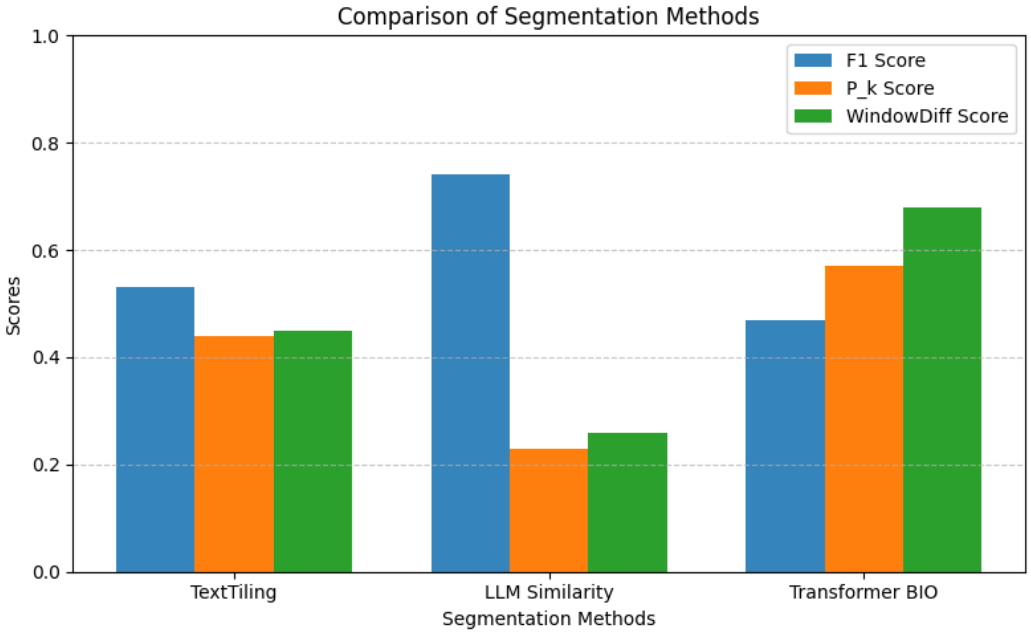

Figure 4.1: Comparative performance visualization of segmentation methods. A higher F1 score (blue) indicates better segmentation accuracy, whereas lower P_k (orange) and WindowDiff (green) scores indicate fewer segmentation errors.

To interpret the results, it is important to note that a higher F1 score indicates better segmentation accuracy, while lower P_k and WindowDiff scores signify improved alignment with the reference annotations. Among the three methods, the LLM Similarity Thresholding approach demonstrates the lowest P_k and WindowDiff scores, suggesting stronger segmentation consistency, whereas the Transformer-based BIO model achieves the highest F1 score, reflecting its ability to capture nuanced topic boundaries effectively.

Ultimately, these findings underscore the importance of integrating both semantic representations and robust labeling protocols to achieve high-fidelity topic segmentation in diverse, unstructured audio transcripts.

4.2.2 Impact of Threshold Variation in LLM-Based Similarity Approach

Beyond the global performance metrics reported in Table 4.1, additional analyses were conducted to investigate how varying the similarity threshold influences the proportion of sentences assigned to topics in the LLM-based method. Figure 4.2 ("Effect of Similarity Threshold on Sentence Assignment Ratio") illustrates this relationship by plotting the average fraction of sentences assigned

as a function of the threshold (ranging from 0 to 1).

At very low thresholds (below 0.2), nearly all sentences in the transcripts surpass the threshold, causing the assignment ratio to approach 1.0. This behavior, while maximizing coverage, can introduce substantial noise and over-segmentation, as even sentences with marginal similarity scores are labeled as belonging to a topic. As the threshold increases, the assignment ratio begins to decline, reflecting a stricter criterion for topic membership. Notably, there is a pronounced inflection near 0.6, beyond which the ratio decreases sharply. Once the threshold exceeds 0.8, only sentences with exceptionally high similarity remain assigned, potentially causing under-segmentation by discarding sentences with moderate topic alignment.

This threshold-dependent pattern underscores a fundamental trade-off in the LLM-based approach:

- **Lower thresholds** facilitate broader inclusion of sentences, mitigating the risk of missing subtle topic cues but risking over-segmentation.
- **Higher thresholds** ensure that only the most confident assignments are made, reducing false positives but risking under-segmentation by excluding sentences with borderline similarity.

In practical terms, an intermediate threshold (e.g., 0.4) often represents a balanced choice, maintaining sufficient coverage of topical content while avoiding excessive misclassification. This balance is further supported by the comparative results in Section 4.2.1, where the LLM-based approach at a 0.4 threshold yielded competitive F1 scores alongside favorable P_k and WindowDiff measures. Consequently, careful calibration of the similarity threshold emerges as a critical factor in achieving robust and reliable topic segmentation performance for podcast transcripts.

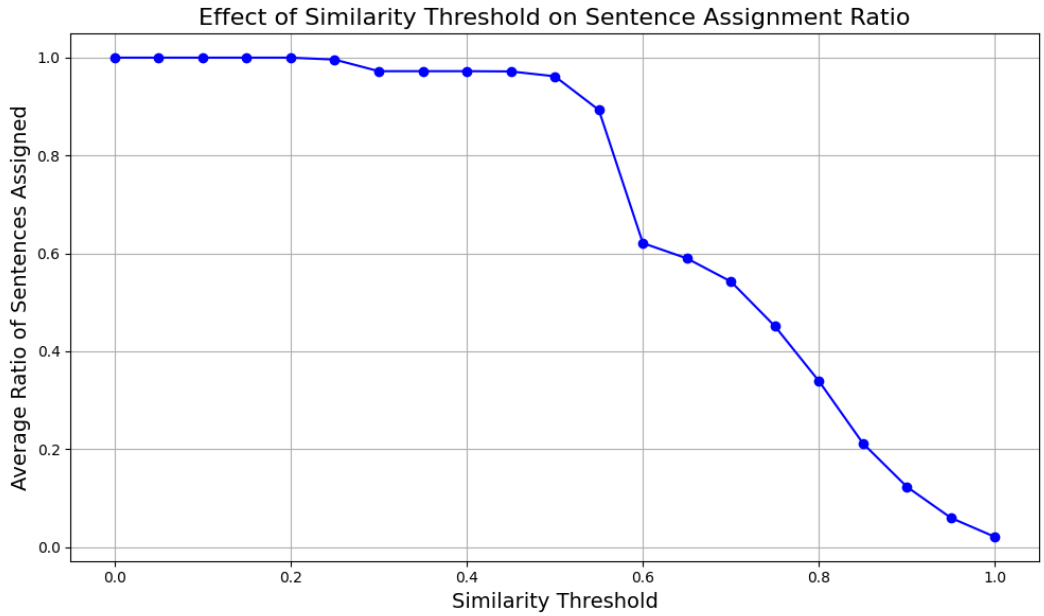

Figure 4.2: Effect of Similarity Threshold on Sentence Assignment Ratio for the LLM-based approach. At lower thresholds, most sentences surpass the threshold and are assigned to topics, whereas at higher thresholds only a minority of sentences qualify, reflecting a trade-off between over-segmentation and under-segmentation.

4.2.3 Threshold Sensitivity Analysis for the LLM-Based Method

To gain deeper insight into the role of threshold calibration in the LLM-based Similarity Thresholding approach, an extended threshold sensitivity analysis was conducted. Figure 4.3 illustrates how each of the three evaluation metrics—F1 Score (blue), P_k (red), and WindowDiff (green)—varies as the similarity threshold increases from 0.0 to 1.0. The following observations can be drawn:

- **Peak in F1 Score Near 0.4:** As the threshold starts at 0.0, many sentences are over-assigned, causing moderate precision and recall. As the threshold approaches 0.4, F1 Score rises to its maximum, indicating a balanced trade-off between detecting most true boundaries and avoiding false positives. Beyond 0.4, F1 Score declines sharply, reflecting the method's increasing tendency to miss boundaries (under-segmentation).
- **Minimum in P_k and WindowDiff Around 0.4:** Both P_k and WindowDiff exhibit their lowest values near the same threshold, signifying

minimal segmentation errors and more precise boundary placement at that point. When the threshold is too low (≤ 0.2), over-segmentation becomes prevalent, inflating error metrics; conversely, when the threshold is too high (≥ 0.8), many legitimate boundaries are overlooked, again increasing error rates.

- **Sharp Changes After the Inflection Point:** All three metrics show relatively stable or gradual changes until approximately 0.6, beyond which F1 Score declines and both P_k and WindowDiff increase significantly. This inflection indicates that once the threshold surpasses a moderate value, the model becomes highly selective, discarding sentences with moderate similarity and thus losing recall.

Hence, the above trends reaffirm that threshold selection is critical for maximizing the LLM-based method's effectiveness. In practice, a threshold in the range of 0.3–0.4 appears to yield the most balanced performance, minimizing over-segmentation at the low end and avoiding excessive under-segmentation at the high end.

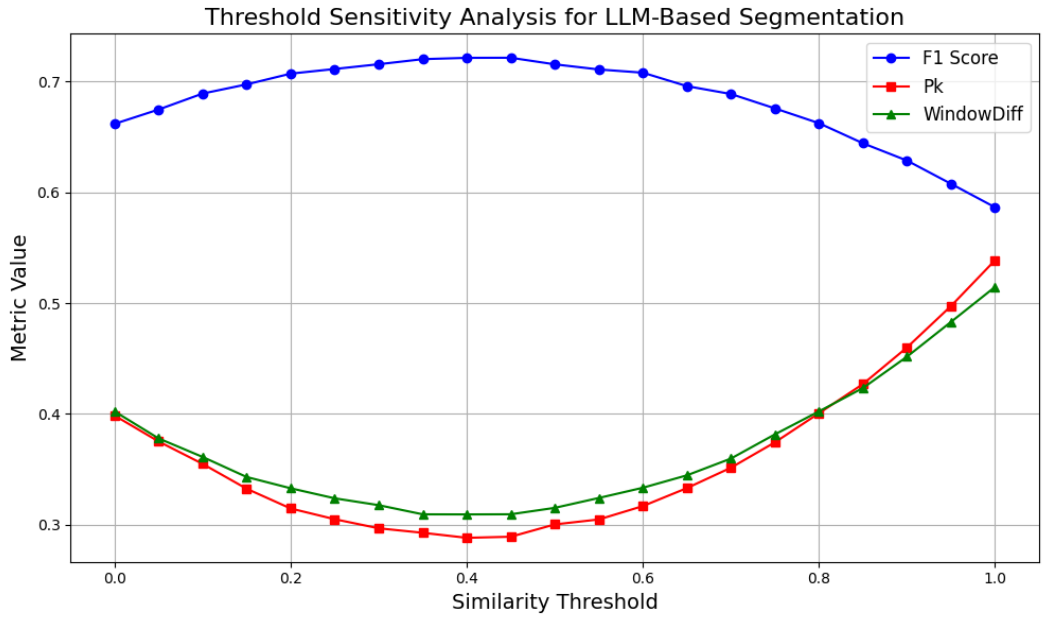

Figure 4.3: Threshold Sensitivity Analysis for the LLM-Based Segmentation. F1 Score peaks around a similarity threshold of 0.4, whereas P_k and WindowDiff are minimized near the same value.

4.2.4 Fold-Wise Performance Analysis for the DistilBERT BIO Method

To gain deeper insight into the variability of the DistilBERT BIO labeling approach, fold-wise F1, P_k, and WindowDiff (WD) scores were examined across all 30 LOOCV folds. Figure 4.4 presents a single visualization capturing the performance trends of all three metrics, illustrating disparities in segmentation quality across different folds.

Performance Trends Across Folds: The analysis highlights substantial variability across folds, emphasizing the challenges posed by different podcast transcripts:

- **F1 Score:** The F1 score fluctuates significantly, ranging from approximately 0.30 to 0.85. Certain folds (e.g., Fold 10 and Fold 28) achieve higher F1 values, indicating successful boundary detection, whereas others (e.g., Fold 7 and Fold 19) fall on the lower end, suggesting difficulties in detecting topic shifts.
- **P_k Score:** The P_k metric, which measures segmentation errors, exhibits notable fluctuations. Some folds, such as Fold 3 and Fold 17, present high P_k values, indicating frequent misalignment between predicted and true boundaries. In contrast, folds like Fold 16 and Fold 28 show lower P_k values, suggesting better segmentation performance.
- **WindowDiff (WD) Score:** The WindowDiff metric follows a similar trend, with certain folds (e.g., Fold 3 and Fold 17) reaching close to 1.0, reinforcing the pattern of boundary misplacement. Meanwhile, midrange WD values in folds like Fold 20 and Fold 25 suggest moderate alignment of predicted and reference boundaries.

Implications and Next Steps: The observed fold-wise variability underscores the heterogeneous nature of the DistilBERT BIO method's segmentation performance. While some transcripts are well-segmented, others introduce significant challenges, resulting in increased segmentation errors. These findings point to several possible directions for improvement:

- **Transcript-Specific Error Analysis:** Reviewing outlier folds qualitatively could reveal if issues stem from unique discourse structures, annotation inconsistencies, or insufficient training data for certain topic shifts.

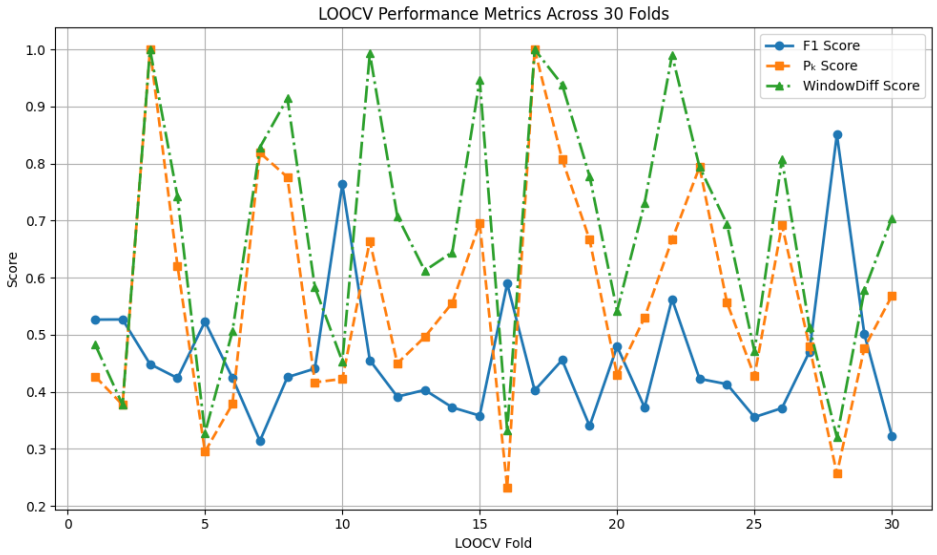

Figure 4.4: Fold-wise Performance Metrics for DistilBERT BIO Labeling, showing F1 Score, P_k, and WindowDiff trends across 30 LOOCV folds. A higher F1 score indicates better segmentation performance, while lower P_k and WindowDiff scores suggest improved boundary alignment.

- **Refinement of Annotation and Labeling Schemes:** Since BIO labeling relies on well-defined boundaries, exploring hierarchical or multi-label tagging schemes may help mitigate ambiguity in complex transcripts.
- **Expansion of Training Data:** Increasing the diversity of annotated transcripts could help the model generalize to a wider variety of speaking styles and conversational formats.

By consolidating the fold-wise F1, P_k, and WindowDiff trends into a unified visualization, this analysis provides a clearer understanding of performance variability. The next section presents a more detailed sentence-level breakdown, examining how well the model differentiates between the `B-`, `I-`, and `O` labels under challenging discourse conditions.

4.2.5 Sentence-Level Confusion Matrix (DistilBERT-BIO)

To evaluate how well the DistilBERT-BIO method distinguishes between topic and non-topic sentences, we analyze the sentence-level confusion matrix shown in Figure 4.5. This analysis provides insights into the model's strengths and

weaknesses in detecting topic boundaries and helps identify potential sources of segmentation errors. The results are categorized into four main classes:

- **True Topic, Predicted Topic (265):** These sentences contain actual topic boundaries and are correctly identified, reflecting the model's ability to recognize sentences that introduce or continue a topic. The high accuracy in this category suggests that the model effectively learns strong topic indicators.
- **True Topic, Predicted Non-Topic (235):** These sentences are truly topic-related but were misclassified as *Non-Topic*. Such errors indicate that the model struggles with subtle topic shifts or shorter topic segments, potentially leading to under-segmentation.
- **True Non-Topic, Predicted Topic (337):** These sentences do not contain any topic boundary but were incorrectly labeled as *Topic*. This suggests that the model sometimes misinterprets transitional or contextual phrases as topic shifts, leading to false positives. Addressing this issue may require additional context-aware filtering.
- **True Non-Topic, Predicted Non-Topic (1163):** The majority of non-topic sentences were correctly classified, highlighting that the model performs well at identifying sentences that do not contribute to a topic transition. This suggests that while false positives exist, the model's baseline ability to distinguish non-topic content is reliable.

Significance of These Results: This confusion matrix analysis provides a detailed understanding of where the model succeeds and where it fails in topic segmentation. The balance between false positives and false negatives is particularly important in refining segmentation performance. While the model reliably detects clear topic transitions, improvements are needed in reducing misclassifications caused by subtle topic cues and ambiguous transitions. These insights inform potential strategies for enhancing the model, such as refining annotation guidelines, incorporating more diverse training data, or leveraging hierarchical segmentation approaches.

Implications for Topic Segmentation

- **Topic Class:** Out of approximaely 500 total topic sentences (265 + 235), the model correctly identifies 265. This yields a recall of roughly 53% and, considering the 337 false positives, a precision of about 44%.

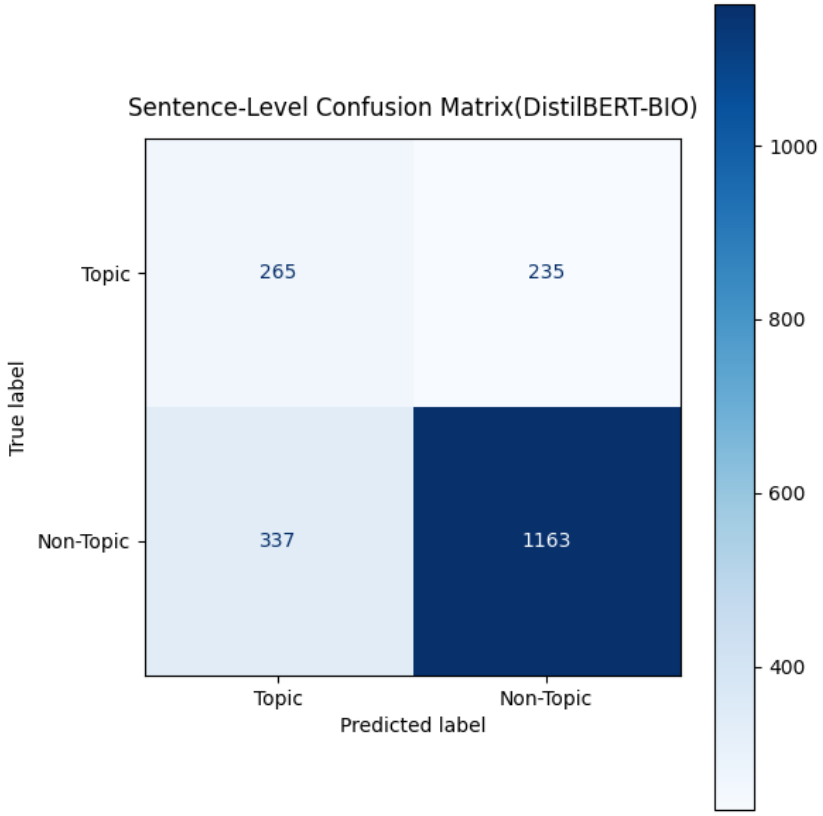

Figure 4.5: Sentence-Level Confusion Matrix (DistilBERT-BIO). Rows represent the true labels, and columns indicate predicted labels.

The resulting F1 score for topic sentences stands near 0.47, highlighting that while the model detects some boundaries effectively, it also struggles with ambiguous or short topic segments.

- **Non-Topic Class:** Out of approximately 1,500 non-topic sentences (337 + 1163), 1163 are correct, indicating a higher success rate for identifying non-boundary content. Nevertheless, the 337 false alarms underscore that the model sometimes confuses discourse transitions with genuine topic shifts.

Potential Refinements.

- **Hierarchical Labeling:** Introducing finer distinctions (e.g., subtopics or minor transitions) might reduce false positives and missed boundaries, especially in multi-speaker or digressive discourse.
- **Data Augmentation:** Increasing the variety of transcripts (particularly those featuring rapid or subtle topic shifts) could help the model learn more nuanced boundary cues.
- **Model Threshold Tuning:** Adjusting decision thresholds or confidence measures might reduce the model's tendency to over-label or under-label topic sentences.

This confusion matrix confirms that while the DistilBERT-BIO method captures a portion of true topic sentences, it also produces moderate levels of both missed topics (false negatives) and false alarms (false positives). Consequently, further strategies such as refined labeling criteria and training data—may be required to bolster sentence-level boundary detection performance.

4.2.6 Fold-by-Fold Performance Variability in DistilBERT BIO Method

Having presented the aggregate metrics for the DistilBERT BIO method, the next step is to examine the variability in performance across individual folds in the Leave-One-Out Cross-Validation (LOOCV) procedure. While some folds demonstrated relatively high F1 scores and moderate P_k/WindowDiff values, others performed poorly, raising questions about the factors influencing such disparities.

To quantitatively analyze this variability, the **standard deviation** and **variance** of each metric were computed. Table 4.2 presents the fold-wise

performance metrics, including the mean, standard deviation, and variance for F1 Score, P_k, and WindowDiff. These additional statistical measures help to highlight the degree of fluctuation in performance across folds.

The observed variability suggests that the **DistilBERT BIO model is highly sensitive to transcript-specific factors**, such as topic coherence, discourse structure, and speaker transitions. Notably, the **WindowDiff score exhibits the highest variance**, indicating that boundary misalignment fluctuates significantly depending on the fold. Conversely, the **F1 score shows relatively lower variance**, implying a more stable predictive capability across different transcripts.

4.2.7 Key Takeaways

- **Highest Variance in WindowDiff:** This suggests that the model struggles with consistently predicting boundaries across different transcripts.
- **Relatively Low Variance in F1 Score:** The model's classification ability remains relatively stable across different folds.
- **Extreme Performance Variability in Certain Folds (e.g., Fold 3, Fold 17):** These folds exhibit outlier behavior with significantly higher P_k and WindowDiff values, indicating poor boundary alignment.

These findings highlight the need for **further refinements** in training strategies, such as additional fine-tuning on diverse podcast datasets, incorporation of domain-specific embeddings, or integration of hierarchical topic modeling techniques.

Analysis of Performance Variability. The substantial variability in F1, P_k, and WindowDiff scores across different folds suggests that transcript characteristics play a crucial role in model performance. The following key factors were identified as influencing segmentation effectiveness:

- **Transcript Complexity:** Transcripts with frequent digressions, informal discussions, or highly unstructured conversations led to lower scores. For example, Folds 3 and 17, which had the highest P_k and WindowDiff values (1.0000), contained transcripts where speakers frequently changed topics abruptly without clear indicators. Such cases made it difficult for the model to learn stable topic boundaries.

Table 4.2: Fold-wise performance metrics for the DistilBERT BIO method across 30 LOOCV folds, along with standard deviation and variance for each metric.

Fold	**F1 Score**	$\mathbf{P_k}$	**WindowDiff**
Fold 1	0.5266	0.4270	0.4831
Fold 2	0.5269	0.3774	0.3774
Fold 3	0.4486	1.0000	1.0000
Fold 4	0.4241	0.6212	0.7424
Fold 5	0.5232	0.2951	0.3279
Fold 6	0.4244	0.3787	0.5059
Fold 7	0.3139	0.8190	0.8286
Fold 8	0.4257	0.7763	0.9145
Fold 9	0.4407	0.4167	0.5833
Fold 10	0.7636	0.4227	0.4536
Fold 11	0.4554	0.6643	0.9930
Fold 12	0.3915	0.4498	0.7081
Fold 13	0.4033	0.4961	0.6124
Fold 14	0.3727	0.5558	0.6447
Fold 15	0.3583	0.6955	0.9465
Fold 16	0.5896	0.2324	0.3320
Fold 17	0.4025	1.0000	1.0000
Fold 18	0.4558	0.8077	0.9385
Fold 19	0.3401	0.6667	0.7778
Fold 20	0.4794	0.4294	0.5418
Fold 21	0.3727	0.5304	0.7304
Fold 22	0.5626	0.6667	0.9902
Fold 23	0.4229	0.7941	0.7941
Fold 24	0.4135	0.5562	0.6938
Fold 25	0.3557	0.4274	0.4710
Fold 26	0.3715	0.6923	0.8077
Fold 27	0.4703	0.4800	0.5133
Fold 28	0.8512	0.2564	0.3205
Fold 29	0.5009	0.4771	0.5780
Fold 30	0.3231	0.5682	0.7045
Average	0.4570	0.5660	0.6772
Std. Dev	0.1160	0.1960	0.2130
Variance	0.0140	0.0380	0.0450

- **Annotation Inconsistencies:** Some transcripts contained inconsistencies in manual annotations, where subtle transitions were labeled differently across transcripts. This impacted model training and resulted in lower F1 scores, particularly in folds where the validation transcript differed significantly in annotation style from the training data.
- **Highly Scripted vs. Spontaneous Speech:** Folds with higher F1 scores (e.g., Fold 10 and Fold 28) corresponded to well-structured transcripts, such as scripted monologues or interview-style podcasts with clear topic transitions. In contrast, folds with lower F1 scores (e.g., Fold 7 and Fold 30) contained free-flowing conversations, where speakers frequently revisited previous topics or digressed, making it harder to segment properly.
- **Topic Overlaps and Ambiguity:** Transcripts in which multiple topics were discussed simultaneously, or where a single segment spanned multiple intertwined topics, tended to increase segmentation errors. For instance, Fold 8 had an F1 score of 0.4257 but a very high WindowDiff (0.9145), suggesting that topic boundaries were ambiguous or difficult to detect precisely.
- **Presence of Repetitive Phrases and Filler Words:** Some podcasts exhibited repetitive speech patterns where the same key phrases were used across different segments. This misled the model into merging or splitting topics incorrectly, increasing segmentation errors.
- **Speaker Variability and Turn-Taking Frequency:** Transcripts with frequent speaker changes caused segmentation confusion. Some folds (e.g., Fold 15 and Fold 19) had lower F1 scores due to multi-speaker conversations where speaker changes were incorrectly inferred as topic boundaries.
- **Insufficient Training Samples for Specific Patterns:** In cases where a specific topic structure (e.g., interview-based transcripts, panel discussions) was underrepresented in training data, folds containing those transcripts performed poorly. This was evident in folds with significantly lower F1 scores and high P_k/WindowDiff values.

Observations from High-Performing Folds:

- **Fold 10 and Fold 28 Achieved the Best Performance:** These folds showed the highest F1 scores (0.7636 and 0.8512, respectively) and the lowest segmentation errors. They contained structured monologues with distinct, well-separated topic transitions, allowing the model to learn clear segmentation patterns.
- **Lower P_k and WindowDiff Scores Indicate Better Alignment:** The lowest P_k values (Fold 16: 0.2324 and Fold 28: 0.2564) corresponded to transcripts where predicted and actual boundaries were well-aligned, suggesting minimal segmentation errors.
- **Balanced Segmentation in Certain Folds:** Some folds (e.g., Fold 5, Fold 16, Fold 27) exhibited an optimal balance of F1, P_k, and WindowDiff, indicating a better match between predicted and gold-standard segmentations. These transcripts featured moderate levels of structure, with transitions neither too ambiguous nor overly rigid.

4.3 Summary

This chapter presented a detailed evaluation of three topic segmentation methods—TextTiling, LLM-based Similarity Thresholding, and Transformer-based BIO Labeling—using a manually annotated podcast dataset as the gold standard. The segmentation performance was assessed through well-established metrics: F1 Score, P_k, and WindowDiff, providing both quantitative and qualitative insights.

The results indicate that the LLM-based Similarity Thresholding method achieved the best overall performance with an F1 Score of 0.72, a P_k score of 0.29, and a WindowDiff of 0.31. This method effectively captured semantic relationships between segments and demonstrated robustness in handling topic boundaries with subtle lexical shifts. However, careful threshold selection was critical to prevent over-segmentation or under-segmentation.

The TextTiling method, serving as a traditional lexical cohesion-based baseline, achieved moderate segmentation quality (F1 Score: 0.53, P_k: 0.44, WindowDiff: 0.45). It was effective in detecting clear topic transitions but struggled with nuanced boundary shifts, especially in conversational podcast transcripts with overlapping lexical cues.

The Transformer-based BIO Labeling method, despite leveraging DistilBERT with a CRF layer for contextual segmentation, exhibited the lowest

segmentation accuracy (F1 Score: 0.47, P_k: 0.57, WindowDiff: 0.68). The supervised nature of the model made it highly dependent on annotation consistency and training data diversity. The performance varied significantly across different folds in the Leave-One-Out Cross-Validation (LOOCV), with certain folds demonstrating strong segmentation ability while others suffered due to factors such as conversational complexity, speaker variation, and ambiguous topic shifts.

The threshold sensitivity analysis for the LLM-based method highlighted that an optimal similarity threshold around 0.4 yielded the best segmentation balance, with lower thresholds causing over-segmentation and higher thresholds leading to excessive under-segmentation. Additionally, the fold-wise analysis of the DistilBERT-BIO method revealed substantial performance variability, with transcripts exhibiting structured monologues achieving high accuracy, whereas those with spontaneous, multi-speaker, or overlapping discussions led to increased segmentation errors.

Further investigations using sentence-level confusion matrices for the Transformer based BIO method provided deeper insights into common segmentation errors. The model demonstrated higher recall in detecting topic transitions but struggled with precision, often misclassifying non-topic sentences as topic boundaries. This highlights the need for improved annotation consistency, hierarchical labeling, and additional training data to enhance performance.

Hence, this evaluation underscores the importance of balancing lexical, semantic, and contextual cues in topic segmentation tasks. While LLM-based approaches showed the most promise, they require careful tuning and further refinement. The findings from this chapter provide valuable directions for improving segmentation models, which are discussed in the next chapter on Conclusions and Future Directions.

Chapter 5

Conclusions

This thesis explored the challenge of topic segmentation in podcast transcripts by developing two advanced methods: an LLM-based similarity thresholding approach and a transformer-based BIO labeling model. These methods were evaluated against the traditional TextTiling baseline to assess their effectiveness in detecting topic boundaries in unstructured, conversational data.

The key findings from this research include:

- The LLM-based similarity thresholding method demonstrated superior performance in balancing topic coherence and flexibility, effectively capturing nuanced topic shifts through semantic embeddings.
- The transformer-based BIO labeling approach, despite requiring labeled data for training, exhibited strong segmentation capabilities by leveraging contextual information through deep learning architectures.
- The baseline TextTiling method, while computationally efficient, struggled to handle complex topic transitions and lacked the semantic understanding necessary for accurate segmentation.
- Error analysis revealed that overly low similarity thresholds led to excessive topic assignments, whereas high thresholds resulted in inadequate segmentation, emphasizing the need for optimal threshold tuning.
- Leave-One-Out Cross-Validation (LOOCV) provided robust validation, highlighting the strengths and weaknesses of each approach across multiple podcast transcripts.

Despite these promising findings, certain limitations were identified:

- The LLM-based method required manual threshold adjustments, which could impact consistency across different datasets.

- The transformer-based model's reliance on labeled data limited its scalability, as manual annotations are time-consuming and costly.
- Some segmentation errors arose due to ambiguous or overlapping topics, indicating a need for improved context modeling.
- The evaluation was conducted on a limited dataset of 30 podcast transcripts, and performance may vary across different genres or audio qualities.

To address these limitations and further advance topic segmentation research, some future directions are proposed:

- **Automated Threshold Optimization:** Developing adaptive thresholding mechanisms using reinforcement learning or dynamic parameter tuning could enhance the robustness of the LLM-based method.
- **Hybrid Models:** Integrating both semantic similarity and sequence labeling techniques may improve segmentation accuracy by leveraging the strengths of both approaches.
- **Weakly Supervised Learning:** Exploring semi-supervised or unsupervised methods to reduce reliance on extensive labeled data could enhance the scalability of transformer-based models.
- **Domain Adaptation:** Investigating how well these methods generalize to different podcast genres or spoken content, such as interviews or educational talks, could provide deeper insights into their applicability.
- **Real-World Deployment:** Implementing the developed models in practical applications, such as podcast indexing, automatic summarization, or search systems, would validate their usefulness in real-world scenarios.

In conclusion, this thesis contributes to the field of topic segmentation by developing and evaluating advanced methods tailored for podcast transcripts. The findings highlight the importance of semantic embeddings and deep learning in improving segmentation accuracy. While challenges remain, the proposed future directions offer a roadmap for refining and expanding upon the presented work, paving the way for more sophisticated and adaptable segmentation models in the future.

Bibliography

Andrew Aquilina, Sean Diacono, Panagiotis Papapetrou, and Maria Movin. An end-to-end workflow using topic segmentation and text summarisation methods for improved podcast comprehension, 2023. URL https://arxiv.org/abs/2307.13394.

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. Sector: A neural model for coherent topic segmentation and classification, 2019. URL https://arxiv.org/abs/1902.04793.

Haitao Bai, Pinghui Wang, Ruofei Zhang, and Zhou Su. Segformer: A topic segmentation model with controllable range of attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12545–12552, Jun. 2023. doi: 10.1609/aaai.v37i11.26477. URL https://ojs.aaai.org/index.php/AAAI/article/view/26477.

Doug Beeferman, Adam L. Berger, and John D. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34:177–210, 1999. URL https://api.semanticscholar.org/CorpusID:2839111.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.

David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 05 2003a. doi: 10.1162/jmlr.2003.3.4-5.993.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003b. ISSN 1532-4435.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz

Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Freddy Y. Y. Choi. Advances in domain independent linear text segmentation, 2000. URL https://arxiv.org/abs/cs/0003083.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 100,000 podcasts: A spoken English document corpus. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.519. URL https://aclanthology.org/2020.coling-main.519/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Jacob Eisenstein and Regina Barzilay. Bayesian unsupervised topic segmentation. In *Conference on Empirical Methods in Natural Language Processing*, 2008. URL https://api.semanticscholar.org/CorpusID:1967279.

Haoyu Gao, Rui Wang, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. Unsupervised dialogue topic segmentation with topic-aware utterance representation, 2023. URL https://arxiv.org/abs/2305.02747.

Azin Ghazimatin, Ekaterina Garmash, Gustavo Penha, Kristen Sheets, Martin Achenbach, Oguz Semerci, Remi Galvez, Marcus Tannenberg, Sahitya Mantravadi, Divya Narayanan, Ofeliya Kalaydzhyan, Douglas Cole, Ben Carterette, Ann Clifton, Paul N. Bennett, Claudia Hauff, and Mounia Lalmas. Podtile: Facilitating podcast episode browsing with auto-generated chapters. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4487–4495. ACM, October 2024. doi: 10.1145/3627673.3680081. URL http://dx.doi.org/10.1145/3627673.3680081.

Iacopo Ghinassi, Lin Wang, Chris Newell, and Matthew Purver. Multimodal topic segmentation of podcast shows with pre-trained neural encoders. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, ICMR '23, page 602–606, New York, NY, USA, 2023. Association

for Computing Machinery. ISBN 9798400701788. doi: 10.1145/3591106.3592270. URL https://doi.org/10.1145/3591106.3592270.

Dimitrios C. Gklezakos, Timothy Misiak, and Diamond Bishop. Treeseg: Hierarchical topic segmentation of large transcripts, 2024. URL https://arxiv.org/abs/2407.12028.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5228–5235, 2004. doi: 10.1073/pnas.0307752101. URL https://www.pnas.org/doi/abs/10.1073/pnas.0307752101.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022. URL https://arxiv.org/abs/2203.05794.

Marti A. Hearst. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997. URL https://aclanthology.org/J97-1003.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. Topic segmentation and labeling in asynchronous conversations. *J. Artif. Int. Res.*, 47(1):521–573, may 2013. ISSN 1076-9757.

Anna Kazantseva and Stan Szpakowicz. Topical segmentation: a study of human performance and a new measure of quality. In Eric Fosler-Lussier, Ellen Riloff, and Srinivas Bangalore, editors, *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–220, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.org/N12-1022/.

Rachna Konigari, Saurabh Ramola, Vijay Vardhan Alluri, and Manish Shrivastava. Topic shift detection for mixed initiative response. In Haizhou Li, Gina-Anne Levow, Zhou Yu, Chitralekha Gupta, Berrak Sisman, Siqi Cai, David Vandyke, Nina Dethlefs, Yan Wu, and Junyi Jessy Li, editors, *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 161–166, Singapore and Online, July 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.sigdial-1.17. URL https://aclanthology.org/2021.sigdial-1.17/.

Jeonghwan Lee, Jiyeong Han, Sunghoon Baek, and Min Song. Topic segmentation model focusing on local context, 2023. URL https://arxiv.org/abs/2301.01935.

Raymond Li, Wen Xiao, Linzi Xing, Lanjun Wang, Gabriel Murray, and Giuseppe Carenini. Human guided exploitation of interpretable attention patterns in summarization and topic segmentation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10189–10204, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.694. URL https://aclanthology.org/2022.emnlp-main.694/.

Jiangyi Lin, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. Multi-granularity prompts for topic shift detection in dialogue, 2023a. URL https://arxiv.org/abs/2305.14006.

Jiangyi Lin, Yaxin Fan, Feng Jiang, Xiaomin Chu, and Peifeng Li. Topic shift detection in chinese dialogues: Corpus and benchmark, 2023b. URL https://arxiv.org/abs/2305.01195.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Kelvin Lo, Yuan Jin, Weicong Tan, Ming Liu, Lan Du, and Wray Buntine. Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3334–3340, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.283. URL https://aclanthology.org/2021.findings-emnlp.283.

Melkamu Abay Mersha, Mesay Gemeda yigezu, and Jugal Kalita. Semantic-driven topic modeling using transformer-based embeddings and clustering algorithms, 2024. URL https://arxiv.org/abs/2410.00134.

Einat Minkov and William W. Cohen. Learning to rank typed graph walks: Local and global approaches. In *Joint Ninth WebKDD and First SNA-KDD Worshop 2007 on Web Mining and Social Network Analysis*, Joint Ninth WebKDD and First SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pages 1–8, 2007. ISBN 9781595938480. doi: 10.1145/1348549.1348550. Joint 9th WebKDD and 1st SNA-KDD Workshop 2007 on Web Mining and Social Network Analysis. Held in conjunction with 13th

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2007 ; Conference date: 12-08-2007 Through 15-08-2007.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Arik Reuter, Anton Thielmann, Christoph Weisser, Benjamin Säfken, and Thomas Kneib. Probabilistic topic modelling with transformer representations, 2024. URL https://arxiv.org/abs/2403.03737.

Alessandro Solbiati, Kevin Heffernan, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Cali. Unsupervised topic segmentation of meetings with bert embeddings, 2021. URL https://arxiv.org/abs/2106.12978.

Kaiqiang Song, Chen Li, Xiaoyang Wang, Dong Yu, and Fei Liu. Towards abstractive grounded summarization of podcast transcripts. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4407–4418, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.302. URL https://aclanthology.org/2022.acl-long.302/.

Majd E. Tannous, Wassim Ramadan, and Mohanad A. Rajab. Tshd: Topic segmentation based on headings detection (case study: Resumes). *Advances in Human-computer Interaction*, 2023:1–12, February 2023. doi: 10.1155/2023/6044007. URL https://downloads.hindawi.com/journals/ahci/2023/6044007.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Jinxiong Xia and Houfeng Wang. A sequence-to-sequence approach with mixed pointers to topic segmentation and segment labeling. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 2683–2693, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599245. URL https://doi.org/10.1145/3580305.3599245.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. TIAGE: A benchmark for topic-shift aware dialog modeling. In

Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1684–1690, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.145. URL https://aclanthology.org/2021.findings-emnlp.145/.

Linzi Xing and Giuseppe Carenini. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 167–177. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.sigdial-1.18.

Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions, 2023. URL https://arxiv.org/abs/2305.10435.

Hai Yu, Chong Deng, Qinglin Zhang, Jiaqing Liu, Qian Chen, and Wen Wang. Improving long document topic segmentation models with enhanced coherence modeling. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5592–5605, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.341. URL https://aclanthology.org/2023.emnlp-main.341/.

Erion Çano and Benjamin Roth. Topic segmentation of research article collections, 2022. URL https://arxiv.org/abs/2205.11249.