Bauhaus-Universität Weimar

Topic Segmentation using Large Language Models

Master's Thesis Defense

Krishna Chaitanya Valaboju
Digital Engineering
Faculty of Media
11.06.2025

First Examiner : Prof. Dr. Benno Stein
Second Examiner : Jun.-Prof. Dr. Jan Ehlers

Supervisor : Dr. Johannes Kiesel

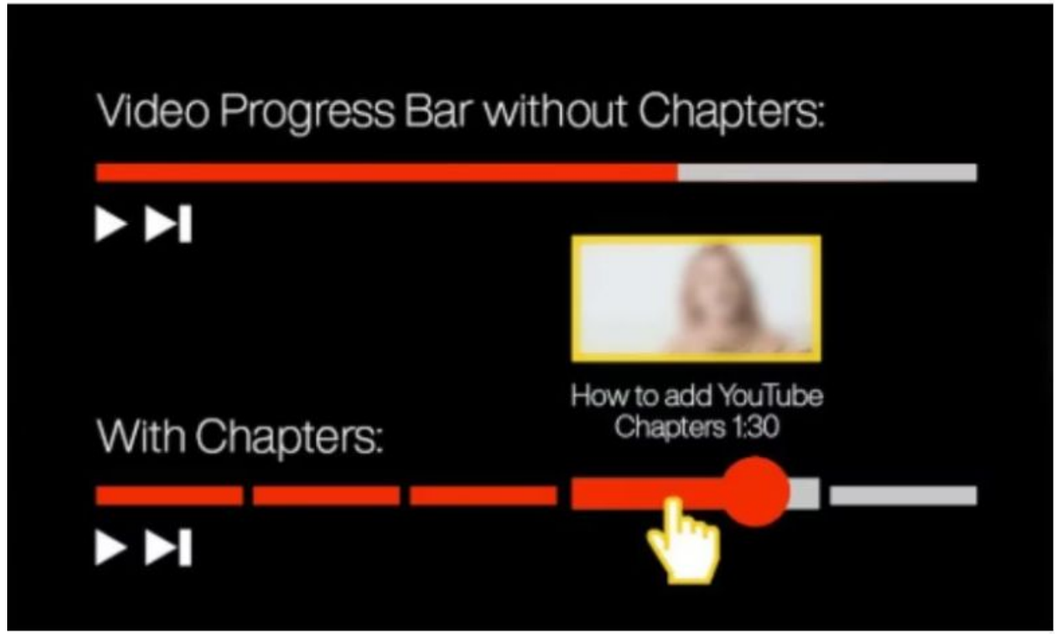
Bauhaus-Universität Weimar

Research Motivation

- Podcasts: 5 million+ episodes, informal & conversational.
- No clear “chapters” = listeners waste time seeking key discussions.
- Unstructured + noisy transcripts make automatic segmentation hard.

YouTube Segmentation Parallel

- Video platforms (e.g., YouTube) already use “chapters” to break long videos into meaningful segments.
- Podcast transcripts: same idea, but only text.
- Goal: Add chapter-like markers to audio via transcripts



YouTube video segmentation with and without chapters, illustrating the difference in the progress bar.

Bauhaus-Universität Weimar

Research Objectives

1. **Build an automated pipeline to split podcast transcripts into topic segments.**

2. **Compare classical (TextTiling) vs. modern methods (LLM-similarity, Transformer BIO).**

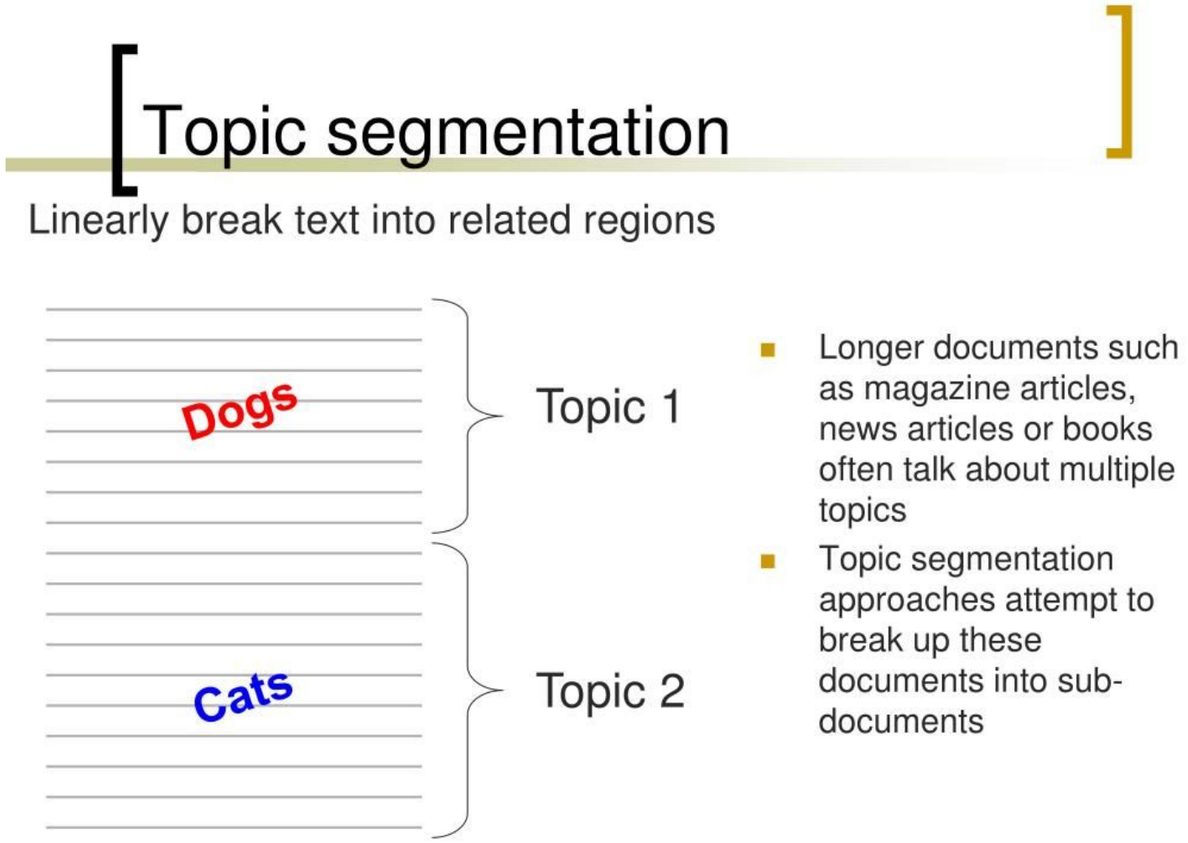
3. **Establish a robust evaluation: manual gold standard, F1, Pk, and WindowDiff.**

4. **Identify the best approach for noisy, conversational data.**

Topic Segmentation: Definition & Importance

What is Topic Segmentation?

- Process: Split a long transcript into coherent segments (distinct topics).
- Why? Transforms unstructured text into manageable, semantically meaningful units.



Bauhaus-Universität Weimar

Applications of Topic Segmentation

Information Retrieval

– Better search & retrieval

Summarization

– Concise, segment-level summaries.

Content Navigation

– Interactive navigation (clickable chapters).

Recommendation Systems

– Improved recommendations (match users to specific segments).

Bauhaus-Universität Weimar

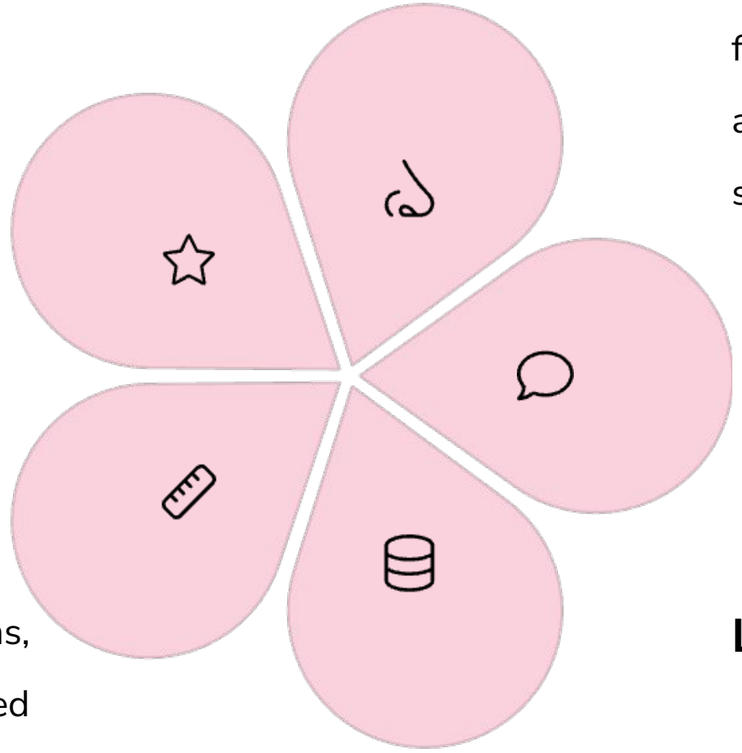
Challenges in Topic Segmentation

Lack of Clear Topic Boundaries

- Podcasts often exhibit informal and conversational speech, leading to ambiguous or overlapping topic boundaries.

Long-form Content

- Podcasts often span extensive durations, posing challenges for models with fixed context windows.



Data Noise

- Transcripts generated via ASR systems frequently contain errors due to factors like accents, background noise, or overlapping speech.

Conversational Nature

- The dynamic and informal nature of podcast conversations adds complexity to segmentation tasks.

Lack of Labeled Data

- The scarcity of annotated podcast datasets hampers the development of robust segmentation models.

Bauhaus-Universität Weimar

Evolution of Topic Segmentation Techniques

Traditional Approaches

TextTiling (Hearst, 1997) and C99 (Choi, 2000) relied on lexical cohesion and fixed sliding windows to detect topic boundaries.

Neural Networks

RNNs and attention-based mechanisms improved capacity to capture long-range dependencies in text.

1 2 3 4

Probabilistic Models

LDA, HMMs, and Bayesian models introduced statistical principles to infer topic boundaries through probabilistic reasoning.

Transformer-Based Models

BERT, RoBERTa, and specialized architectures like Longformer revolutionized segmentation with self-attention mechanisms.

Bauhaus-Universität Weimar

Limitations of Traditional Rule-Based Methods

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.

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.

Dynamic Vocabulary

In conversational domains like podcast transcripts, speakers often employ colloquial language, domain-specific terminology, filler words, or abrupt changes in style. Rule-based methods relying on predetermined lexical cues cannot easily adapt to these variations.

Bauhaus-Universität Weimar

Methodology Overview

Creating a Dataset for Topic Segmentation

Collection of podcast transcripts, preprocessing, and filtering to form a suitable dataset for analysis

Manual Annotation of Topic Segments

Development of annotation guidelines and execution of manual labeling to create a gold standard

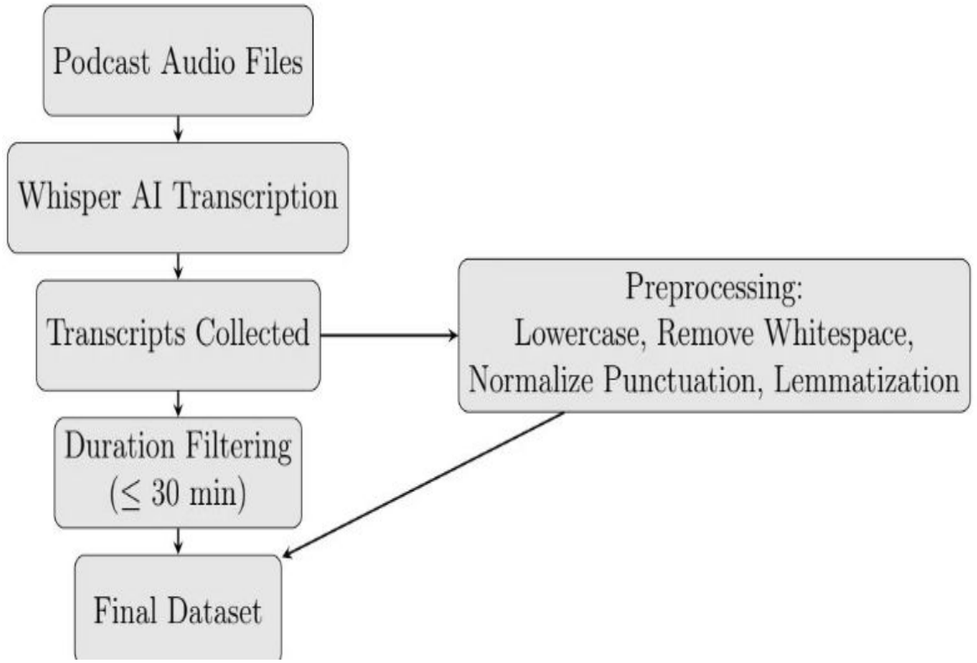
Implementation of Automated Segmentation Methods

Development and optimization of TextTiling, LLM-based, and Transformer-based approaches

Evaluation and Comparative Analysis

Assessment of segmentation performance using established metrics and detailed error analysis

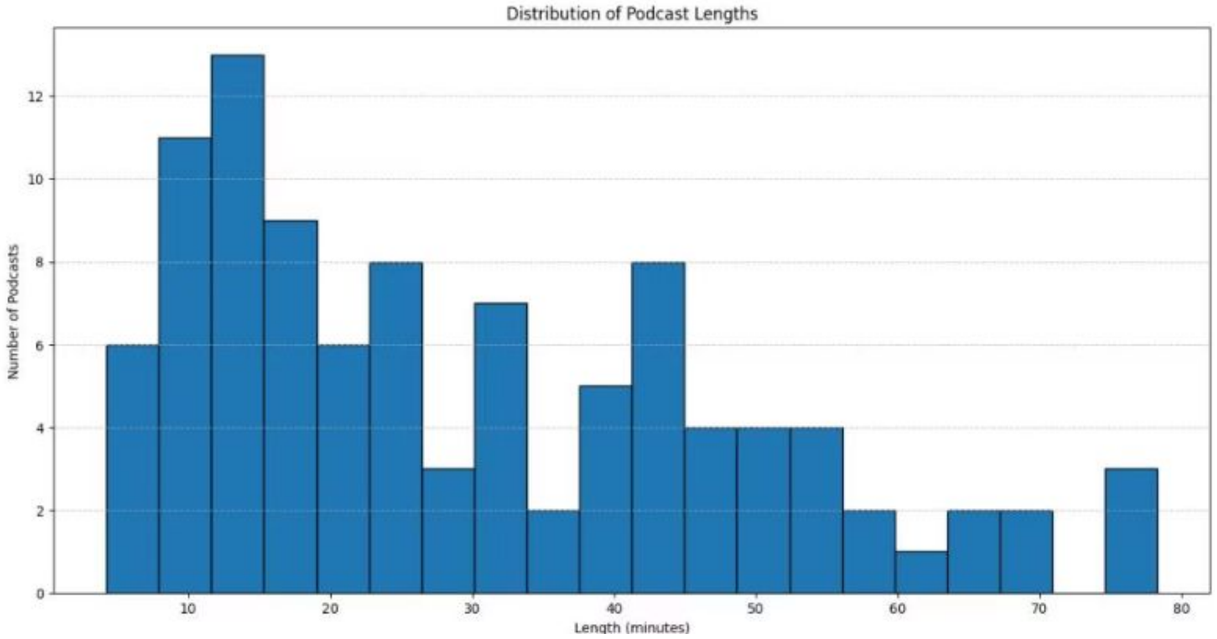
Data Collection and Preprocessing



Process Details

- Collect 100 podcast audios → transcribe via Whisper ASR.
- Filter: keep 30 episodes (10–30 min)
- Minimal Cleaning: lowercase, normalize punctuation, and lemmatize.

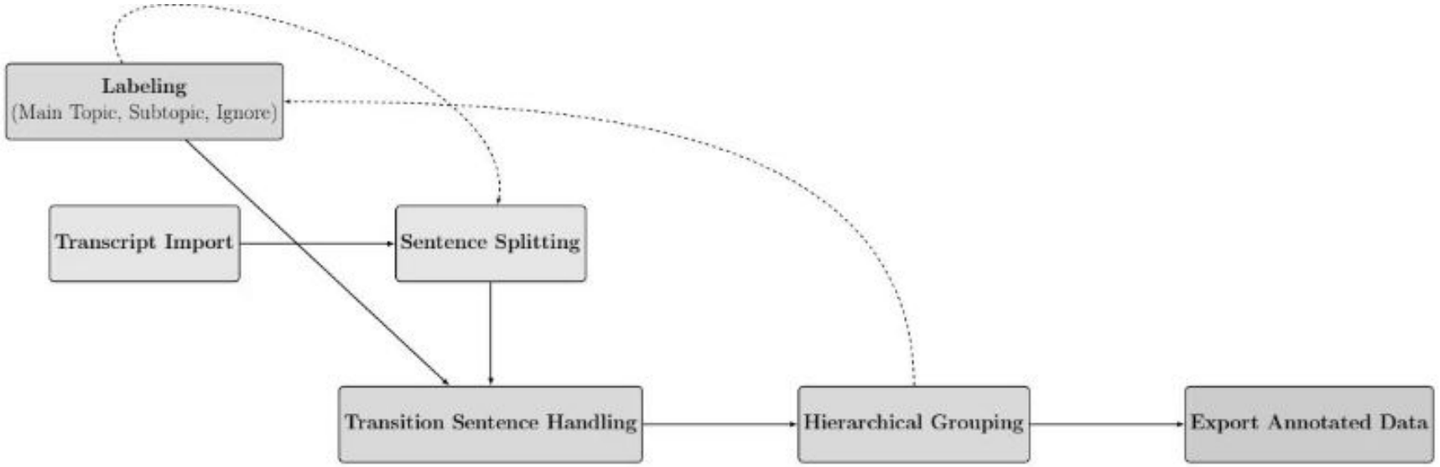
Podcast Duration Distribution



Duration Analysis

- (30 episodes, 10–30 min range)

Manual Annotation Workflow



Annotation Process

- Sentence segmentation → 1–5 sentence spans.
- Hierarchical labels:
  - Main Topic
  - Subtopic
  - Ignore
- Overlap: transition sentences tagged in both segments

Bauhaus-Universität Weimar

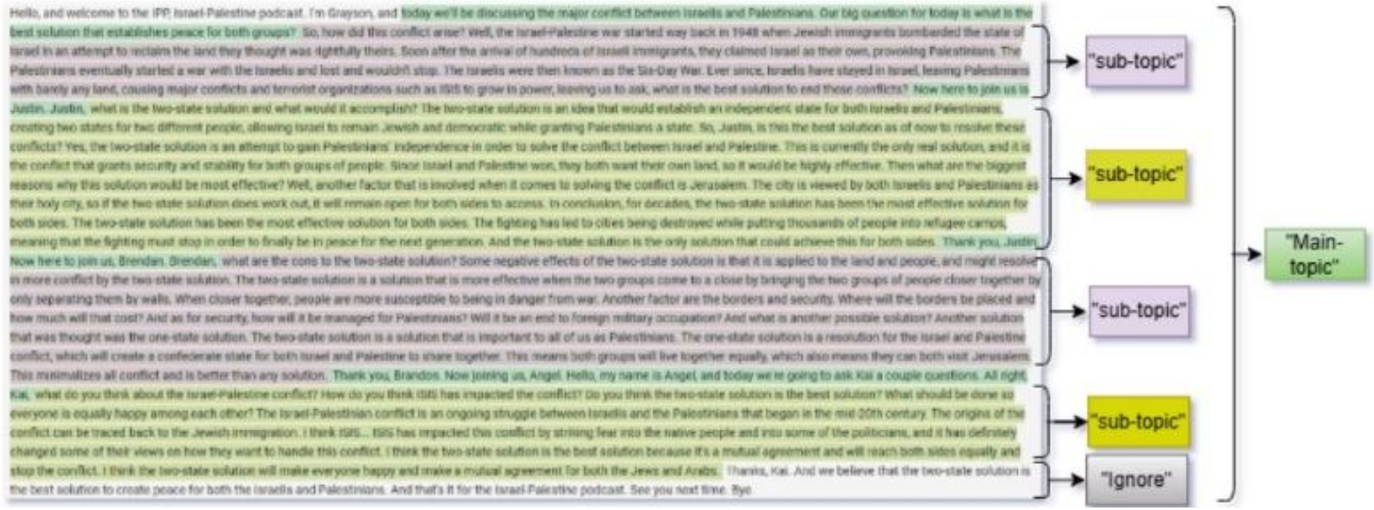
Annotation Example

Hierarchical Labeling Scheme (color-coded)

- *Main Topic :* primary segment

  Eg: (Green segment in the figure)
- *Subtopics:* secondary segment

  Eg: (purple and yellow segments in figure)
- *Ignore*: Refers to content that does not contribute to topic segmentation



Note: Sentences that serve as bridges between topics are included in both adjacent segments

Bauhaus-Universität Weimar

Segmentation Methods

TextTiling (Baseline)

– Lexical cohesion via fixed sliding windows.

– Detects sharp word-frequency dips as boundaries.

LLM-Based Topic Extraction with Similarity Thresholding method

– Use an LLM to propose topic labels & mpnet embeddings.

– Assign sentences when cosine similarity > θ.

Transformer-Based BIO Labeling method

– DistilBERT + CRF fine-tuned on BIO labels (B/I/O).

– Sentences tagged “Boundary” vs. “Inside” vs. “Ignore.”

TextTiling (Baseline) Implementation

Implementation Workflow

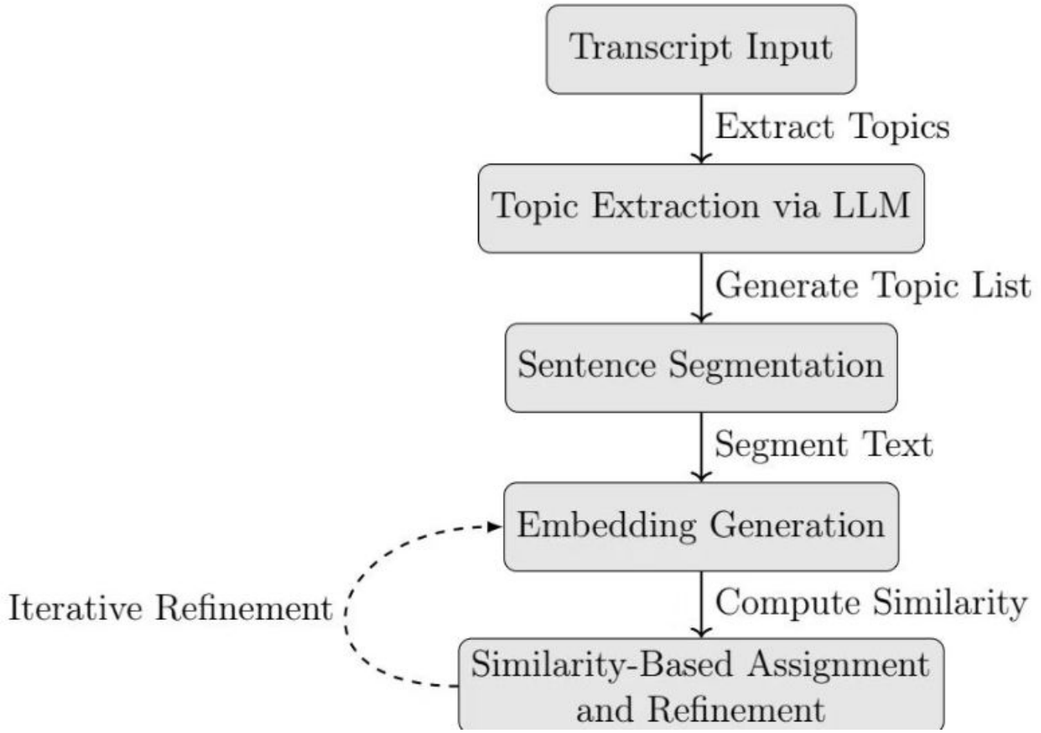
1. Pre-split: Insert breaks every 5 sentences.
2. Tokenize: Split transcript into sentences.
3. Slide windows of 5 sentences: compute lexical cosine similarity.
4. Boundary = lowest similarity point.

House of Wax is a 1953 American warnercolor 3-D horror film about a disfigured sculptor who repopulates his destroyed wax museum by murdering people and using their wax-coated corpses as displays. Directed by Andre DeToth and starring Vincent Price, it is a remake of Warner Bros.' "Mystery of the Wax Museum" (1933), without the comic relief featured in the earlier film."House of Wax" was the first color 3-D feature from a major American studio and premiered just two days after the Columbia Pictures film "Man in the Dark", the first major-studio black-and-white 3-D feature.It was also the first 3-D film with stereophonic sound to be presented in a regular theater.
It premiered nationwide on April 10, 1953 and went out for a general release on April 25, 1953.In 1971, it was widely re-released to theaters in 3-D, with a full advertising campaign.Newly-struck prints of the film in Chris Condon's single-strip StereoVision 3-D format were used.
Another major re-release occurred during the 3-D boom of the early 1980s.In 2005, Warner Bros. distributed a new film also called "House of Wax", but its plot is very different from the one used in the two earlier films.The film starred Elisha Cuthbert, Chad Michael Murray, Paris Hilton and Jared Padalecki. This version received largely negative reviews from critics. In 2014, the film was deemed "culturally, historically, or aesthetically significant" by the Library of Congress and selected for preservation in the National Film Registry.

LLM-Based Topic Extraction with Similarity Thresholding

Implementation Workflow

1. Ask LLM for ≤5 topic labels
2. Split the transcript into sentences.
3. Compute mpnet embeddings for each sentence & each topic.
4. Assign sentences to a topic when cos(sim) > θ (θ=0.4).
5. Iteratively reassign to refine boundaries.

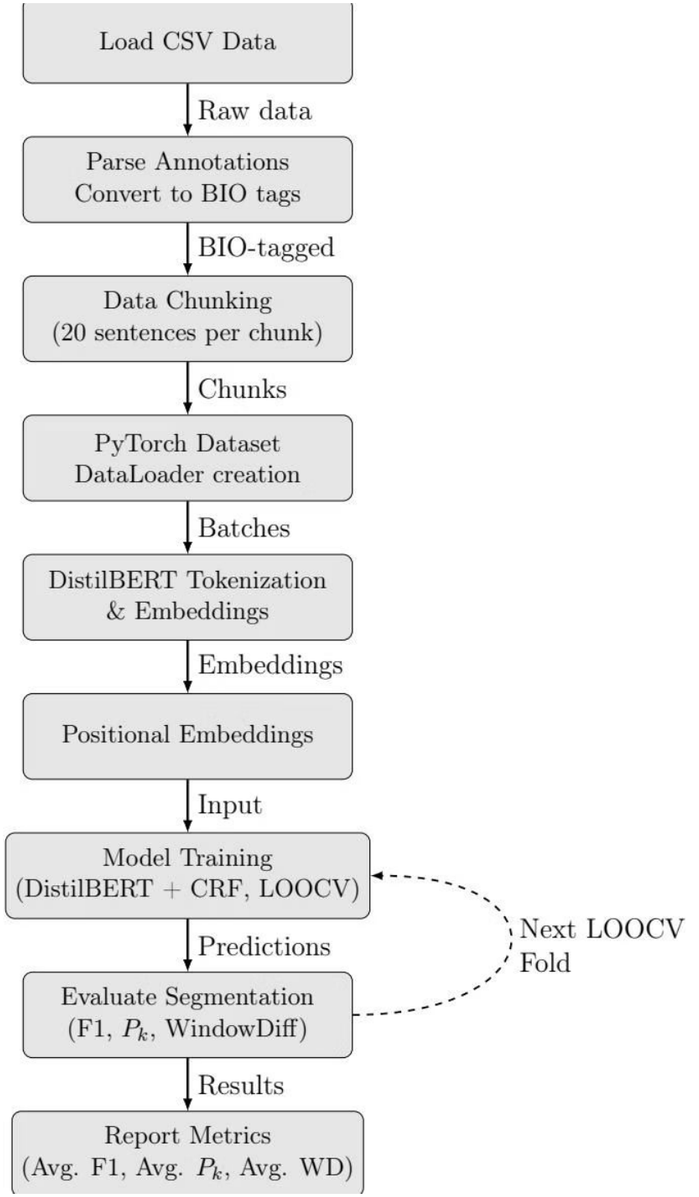


Transformer-Based Model with BIO Labeling

Bauhaus-Universität Weimar

Implementation Workflow

1. Manual BIO tags: label each sentence B/I/O.
2. Chunk transcripts: 20 sent./chunk → max 512 tokens.
3. Tokenize with DistilBERT + positional encodings.
4. Add CRF on top for sequence tag decoding.
5. Train via Leave-One-Out CV (30 folds).
6. Evaluate using F1, Pk, and WindowDiff.



Bauhaus-Universität Weimar

Evaluation Metrics

F1 Score

F1 Score = 2 × (Precision × Recall)/(Precision + Recall)

– Measures correct vs. missed boundaries.

Pk Metric

– Probability two sentences k apart are missegmented.

–A lower Pk score indicates better performance, demonstrating fewer segmentation errors.

WindowDiff Metric

- Counts boundary count differences in a sliding window.

– More sensitive to near-missed boundaries than Pk.

- Similar to Pk, a lower WindowDiff score signifies better segmentation.

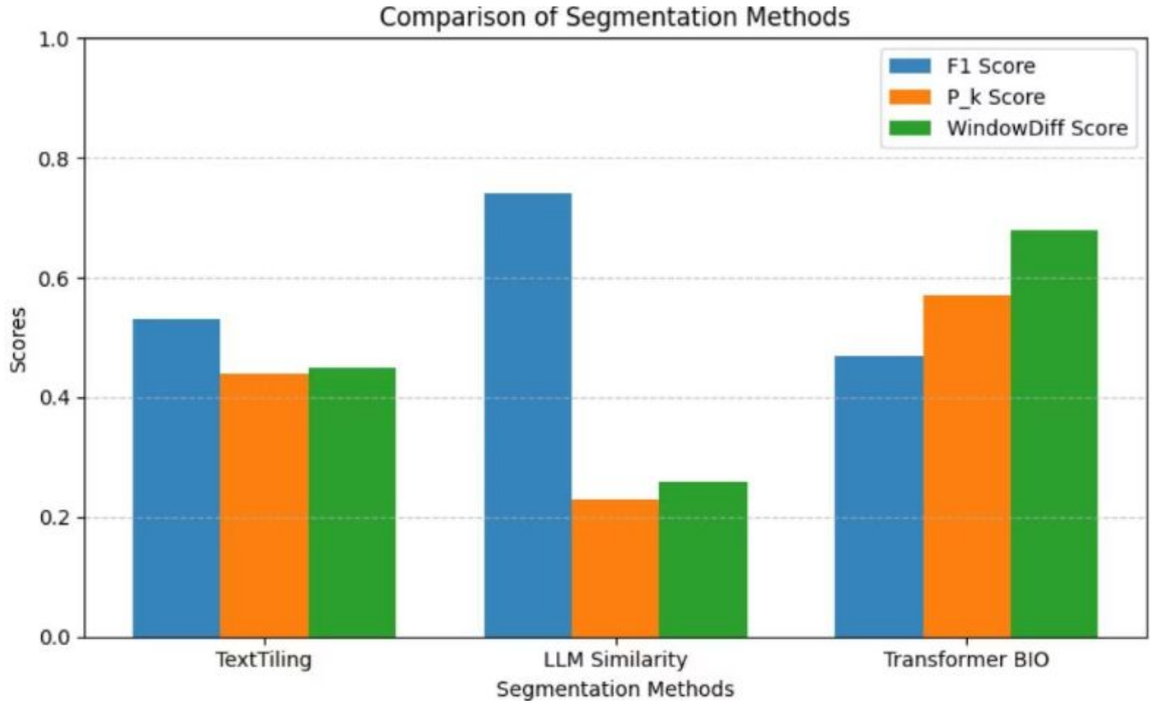
Bauhaus-Universität Weimar

Quantitative Results

Method	F1 Score	Pk	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

- **Best = LLM Similarity Thresholding method** (highest F1, lowest Pk & WD)

Comparative Performance Visualization



Performance Analysis

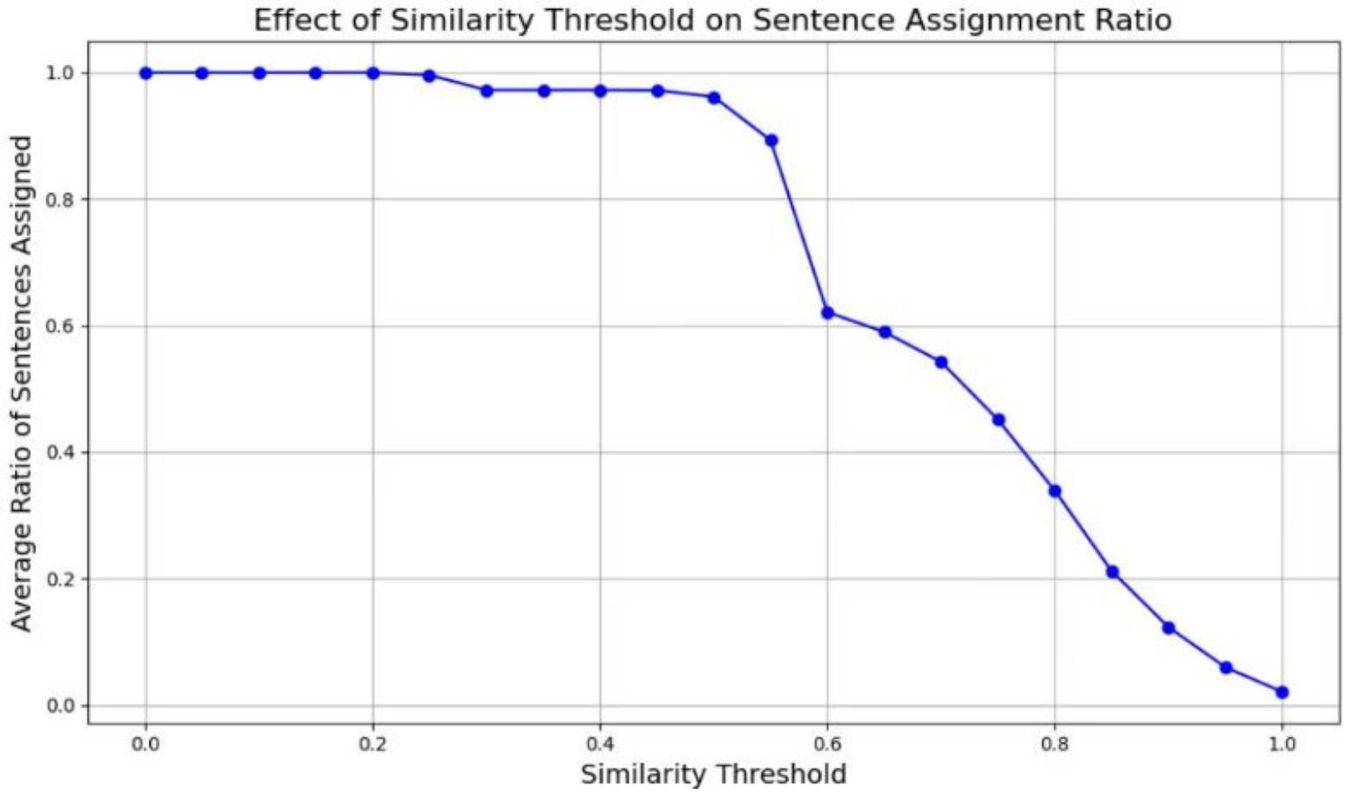
- The LLM-based Similarity Thresholding method substantially outperforms TextTiling on all reported metrics (F1 score of 0.72, Pk of 0.29, and WindowDiff of 0.31).
- Its reliance on contextual embeddings rather than raw lexical overlap allows it to better handle semantically subtle boundaries, such as transitions involving synonyms or related concepts.
- The Transformer-based BIO Labeling approach, which achieved an F1 score of 0.47, a Pk of 0.57, and a WindowDiff of 0.68, performs variably in comparison to TextTiling.

Bauhaus-Universität Weimar

Threshold Sensitivity in LLM-Based Method

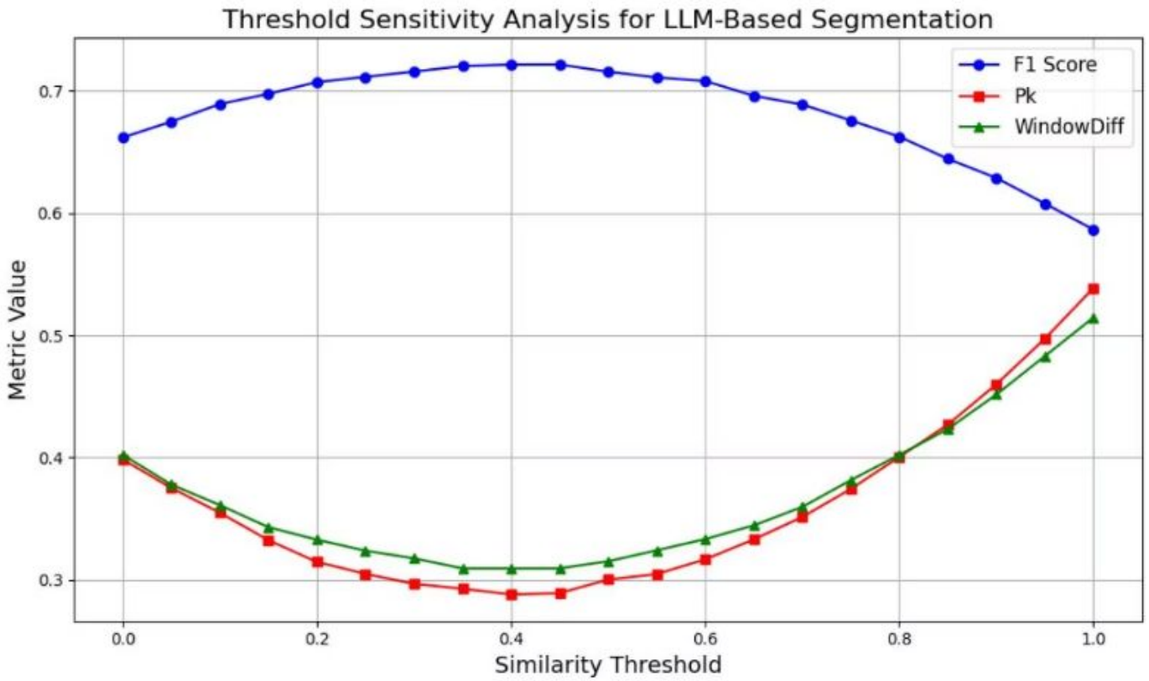
Impact of Threshold Variation

- F1 peaks at θ ≈ 0.4
- Pk & WD minimized at θ ≈ 0.4



Bauhaus-Universität Weimar

Threshold Sensitivity Analysis



Key Observations

- Peak in F1 Score Near 0.4
- Minimum in Pk and WindowDiff Around 0.4
- Sharp Changes After the Inflection Point
- A threshold in the range of 0.3–0.4 appears to yield the most balanced performance.

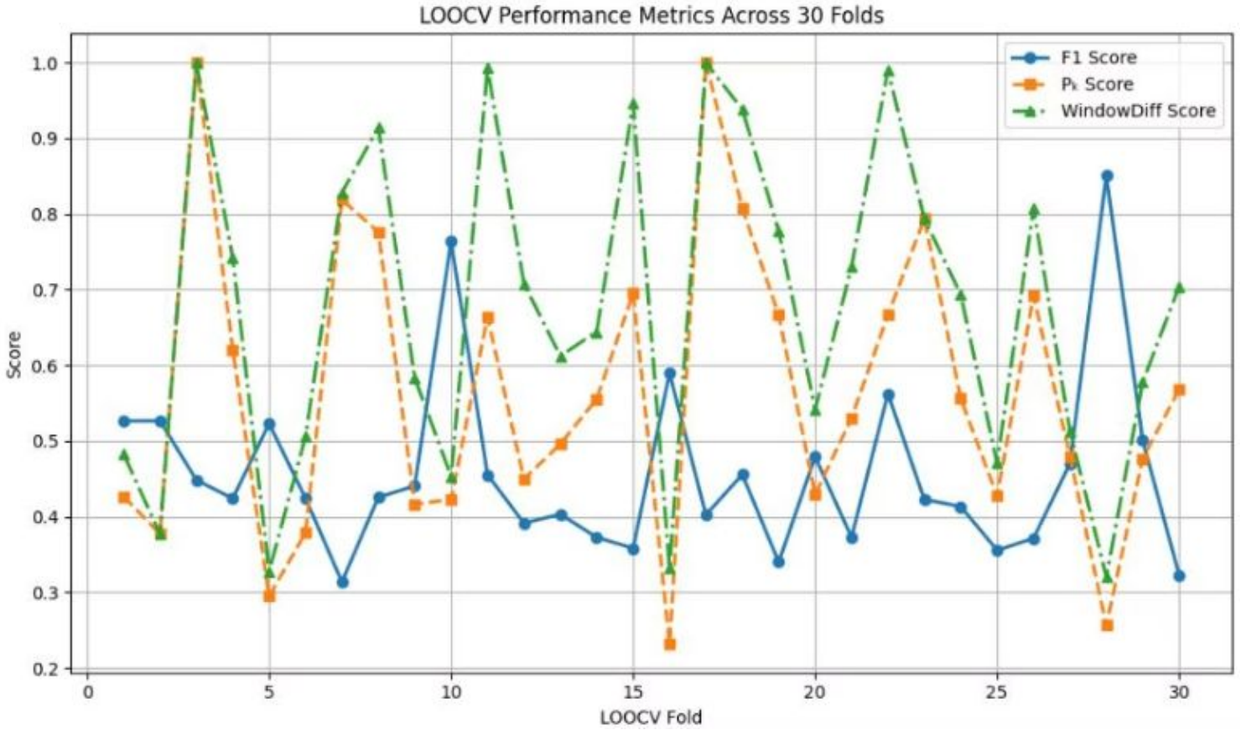
Bauhaus-Universität Weimar

Transformer BIO: Fold-Wise Performance (LOOCV)

Performance Trends Across Folds

- F1 varies from 0.30 to 0.85 across 30 folds
- Pk errors peak in some folds (1.0)
- WD errors also fluctuate widely.

Model highly sensitive to transcript structure & annotation.



Bauhaus-Universität Weimar

Factors Influencing Segmentation Performance

Transcript Complexity

Transcripts with frequent digressions, informal discussions, or highly unstructured conversations led to lower scores. Such cases made it difficult for the model to learn stable topic boundaries.

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.

Scripted vs. Spontaneous Speech

Folds with higher F1 scores corresponded to well-structured transcripts, such as scripted monologues or interview-style podcasts with clear topic transitions.

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.

Key Findings

LLM-based method superiority

The LLM-based similarity thresholding method demonstrated superior performance in balancing topic coherence and flexibility

Threshold sensitivity

Optimal threshold selection (0.3-0.4) is critical for balancing over-segmentation and under-segmentation

Transcript structure impact

Well-structured transcripts with clear transitions yielded better segmentation results

Contextual understanding

Deep semantic representations outperformed surface-level lexical features

Future Directions

Automated Threshold Optimization

Develop adaptive thresholding mechanisms using reinforcement learning

Hybrid Models

Integrate semantic similarity and sequence labeling techniques

Weakly Supervised Learning

Explore semi-supervised methods to reduce reliance on labeled data

Domain Adaptation

test on various podcast genres

Real-World Deployment

Deploy in live podcast apps (chaptering & search)

Questions?

Thank you for your attention!