Explainable Topic Continuity in Political Discourse: A Sentence Pair BERT Model Analysis

Juan-Francisco Reyes *pacoreyes@protonmail.com*
Institute of Computer Science
Brandenburgische Technische Universität Cottbus-Senftenberg
Cottbus, 03046, Germany

Abstract

This study leverages *Sentence Pair Modeling* (*SPM*), *BERT*, and the *Transformers Interpret* library to analyze topic continuity in political discourse. Defined by specific linguistic features, topic continuity is crucial for understanding political communications. Using a dataset of 2,884 sentence pairs, we fine-tuned *TopicContinuityBERT* to focus on how these linguistic features influence topic continuity across sentences. Our analysis reveals that coreferentiality, lexical cohesion, and transitional cohesion are pivotal in maintaining thematic consistency through sentence pairs. This research enhances our understanding of political rhetoric and improves transparency in *natural language processing* (*NLP*) models, offering insights into the dynamics of political discourse.

Keywords: Topic Continuity, Text Segmentation, Sentence Pair Modeling, Explainable AI, BERT, Transformers Interpret.

1. INTRODUCTION

By employing SPM techniques, our research analyzes linguistic features that indicate topic continuity between two sentences in American politics, which may enhance the understanding of political rhetoric in the English language. We treat *topic continuity* in this study as the presence of specific linguistic markers that suggest a sustained subject or theme between two consecutive sentences. Although the term SPM was used for the first time in 2016 by Yin et al., the first study of the meaning across sentence pairs can be found in the research in psycholinguistics of Haviland and Clark (1974), in the context of understanding human processing and comprehension of text. Later, in the 1990s, other studies initiated the use of computational resources to linguistically analyze sentence pairs (Gale & Church, 1993; Haruno & Yamazaki, 1996; Melamed, 1990).

Topic continuity significantly influences the structure and interpretation of conversations, especially within the complex field of political communication (Givón et al., 1983; Fletcher, 1984; Anjali M. & Babu Anto, 2014). What defines this continuity is crucial for understanding political narratives and their impact on public discourse. The inherent ambiguity of political language, characterized by its strategic rhetoric and stylistic complexities (Chomsky, 1988; Orwell, 1946), poses significant challenges to computational models developed for parsing and interpreting such texts. This research, therefore, does not attempt to establish a novel model, improve empirical performance, or introduce a dataset for comprehensive text segmentation. Instead, it focuses on a detailed examination of five linguistic features that define topic continuity between two consecutive sentences: *coreferentiality*, *lexical cohesion*, *semantic cohesion*, *syntactic parallelism*, and *transitional cohesion.*

Although frequently assessing topic continuity requires a nuanced approach that transcends the simple classification of continuity between sentence pairs, the creation of a binary topic continuity dataset, grounded in the methodology of SPM, provides a foundational basis for exploring these shifts, especially with limited data. Using the capacity of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) to handle sentence pairs for sequence classification and the capacity of the Transformers Interpret library (Pierse, 2024) to explain the linguistic features that connect or separate sentence pairs with high granularity, this study aims to understand the linguistic features that define the continuity of a topic, and implicitly its boundary.

After crafting a linguistic-aware rule-based model used for feature extraction, we define it as a baseline for comparison against more sophisticated models while providing a unique integration of linguistics with explainable AI to dissect and understand the subtleties and complexities of political discourse. Such integration offers insightful perspectives on how *machine learning* (*ML*), particularly through models like BERT, can aid in elucidating the nuanced dynamics of topic continuity in political communication. Leveraging the BERT model's performance, an extensive analysis using AI explainability techniques is conducted. This analysis is vital for enhancing model transparency and accountability in NLP, particularly in politics, where language is often used ambiguously.

This paper addresses two pivotal research questions:

- *RQ1*: Which linguistic features contribute to predicting topic continuity in political discourse?
- *RQ2*: How do explainability measurements from the Transformers Interpret library quantitatively and qualitatively relate to the classification of topic continuity features in sentence pairs?

The contribution of this work is three-fold: (1) a balanced dataset of 2,884 pairs of sentences capturing the dynamic nature of topic continuity in political discourse; (2) a systematic analysis of core linguistic features that define topic continuity; and (3) an AI explainability analysis of a BERT model for topic continuity detection using Transformers Interpret to understand the complexities of processing political language.

The remainder of this paper is organized as follows. Section 2 reviews the Related Work, highlighting the main studies on topic continuity and sentence pair modeling in NLP. Section 3 details the Models, including dataset construction, model architecture, training procedure, and experimental setup. Section 4 presents the Results and Discussion, reporting the model's performance metrics and interpretability findings. Finally, Section 5 draws the Conclusion, summarizing the main contributions and suggesting directions for future work.

2. RELATED WORK

The study of topic continuity is deeply related to several areas within NLP, such as text segmentation, topic segmentation, topic change detection, discourse segmentation, text tiling, text chunking, or topic boundary detection (Fan et al., 2024; Pi et al., 2024; Wang et al., 2023). These research niches collectively explore aspects of cohesion and coherence, which are crucial aspects for maintaining a seamless flow of topics within a text, ensuring that the information presented is logically and semantically interconnected (Abdolahi & Zahedi, 2016; Carrell, 1982).

Given the complex nature of discourse, topic continuity is highly interlaced with topic boundary detection; hence, it is a multi-dimensional issue in topic management (Drew & Heritage, 1992; Schiffrin, 1994; Sidnell, 2010; Tannen, 1984), and its study cannot be reduced to the classifying

of sentence pairs. For instance, during a discourse, topics frequently divert temporally (*digression*) to return afterward to the main topic. Consider the following passage:

> "[1] *How can you be against that?*" [2] *And the other side is going around trying to make me sound extreme like I'm an extremist.* [3] *I'm not against that.*"

In this passage, the first and third sentences clearly address the same subject—subtlety supporting "*that*"— while the second sentence diverts temporarily to another subject, illustrating the typical challenge of modeling topic continuity as a mere binary classification of sentence pairs. Nevertheless, studies leverage sentence pairs to study the linguistic features that define topic continuity. For example, Davison (1984) analyzed sentence pairs to examine how syntactic and semantic properties signal topic continuity, focusing on how certain linguistic features mark sentence topics and maintain coherence across discourse. Likewise, Greenspan and Segal (1984) use sentence pairs to study the mechanisms that relate a sentence to its nonlinguistic environment and those that relate a sentence to its linguistic context. Fletcher (1984) presented experiments where two short sentences were combined into one, finding that the form of the referent in the second sentence depended on its continuity with the topic of the first sentence, highlighting the use of unmarked linguistic features in cases of high topic continuity.

In the era of ML dominance, Newman et al. (2005) used a decision tree classifier for recognizing textual entailment and semantic equivalence between sentence pairs using linguistic features, and Zhao et al. (2015) used word embeddings and traditional linguistic features in sentence pair classification, demonstrating that combining these features improves performance in textual entailment and semantic relatedness. More recently, the SPM method has been more widely employed in the study of NLP tasks involving sentence pairs because it adhered to simplifying complex discourses into manageable decisions and mapping sentence pairs to representations that capture their semantic relationships (Yang et al., 2023; Yu et al., 2019). By focusing on whether a sentence continues the topic or indicates a shift, SPM facilitates clearer segmentation, contributing to model interpretability, as it offers discrete, clear conclusions that are easier to analyze and understand (Peng et al., 2023; Yin et al., 2016). In 2016, Yin et al. used attention-based convolutional neural networks (ABCNN) to study if one sentence logically follows from another in the task of selecting the most relevant answer from a pool of candidate answers for a given question; hence, this study can be considered an early antecedent of explainability in NLP using SPM. Subsequent studies have applied SPM to diverse tasks, such as enhancing BERT's performance through transfer fine-tuning with phrasal paraphrases (Arase et al., 2021); measuring general similarity (Shen et al., 2017); reviewing academic papers based on their titles and abstracts (Duan et al., 2019); exploring explainability in CNNs using attention mechanisms (Xu et al., 2020); and mapping relationships between devices with Internet of Things (IoT) technology (Yu et al., 2021). However, there is a gap in the study of traditional linguistic features that define whether sentence pairs define topic continuity or not using SPM to analyze political discourse.

By traditional linguistic features in topic continuity, we mean lexical, morphosyntactic, and discourse-level cues that maintain continuity in a discourse, and not abstract representations of language as they are most frequently used in ML, such as *N-grams*, *word embeddings*, *Term Frequency-Inverse Document Frequency* (*TF-IDF*), *bag of words*, Etc. The traditional linguistic markers that define the continuity of a topic have been the focus of classic and modern researchers in linguistics. For instance, Ariel (1990) introduced the idea that pronominalization (coreferentiality), or the use of pronouns, can indicate continuity if they refer back to entities mentioned in previous sentences or signal a shift if new referents are introduced without clear antecedents. The taxonomies developed by Halliday and Hasan (1976) emphasized lexical cohesion, highlighting that the presence or absence of lexical ties between sentences, such as repetition, synonyms, or related terms, helps maintain topic continuity, while a sudden drop in

lexical cohesion might signal a topic shift. Givón (1995) noted that syntactic parallelism, or the use of similar sentence structures, often indicates topic continuity, whereas a change in sentence structure might suggest a topic shift. Van Dijk (1980) explored semantic cohesion, suggesting that changes in the semantic field or theme from one sentence to another can mark a topic shift, such as shifting from discussing a historical event to detailing a personal anecdote. Halliday and Hasan (1976) and Halliday and Matthiessen (2014) discussed transitional cohesion, where the use of conjunctions and transitional phrases ("*and*", "*however*", "*but*") can either show a continuation of a topic or introduce a contrast or shift, with the absence of such connectives possibly indicating a more abrupt topic shift.

More recently, studies using corpora in the English language support how the linguistic features of interest in this research contribute to the dynamics of topic continuity. Abdalla et al. (2023) explored linguistic features contributing to semantic relatedness, including coreferenciality. Ryu and Jeon (2023) analyzed coreferenciality and semantic cohesion in high school English reading texts using the Coh-Metrix tool (Graesser et al., 2004). On the other hand, Tang and Moindjie (2024) compared human and automatic-generated translations and found that lexical cohesion plays a vital role in topic continuity across translations. Similarly, Batubara et al. (2021) examined lexical cohesion in newspaper discourse and showed that repetition and synonymy dominate in maintaining thematic flow in English news articles. Finally, Putri and Sudaryat (2021) studied parallel structures as part of grammatical cohesion in a Sundanese novel translated into English, finding various parallel sentence patterns contributing to text clarity and cohesion.

In general, we found a gap in the research on the explainability of *large-language models* (*LLMs*) through SPM, but moreover, we did not find a granular study of the linguistic features of topic continuity using pairs of sentences in an LLM. In this sense, it is unique and opens a new path in the area of text segmentation.

3. MODELS

This section presents the model architecture and training process (Section 3.1). We detail the design and implementation of TopicContinuityBERT and its prototypes, including BERT1 and BERT2. We describe the data preparation procedures, model training, and hyperparameter optimization. Additionally, we outline the evaluation metrics and experimental setups employed for assessing model performance (Section 3.2).

In this study, we introduce (1) *TopicContinuity*, a dataset of 2,884 sentence pairs, and (2) *TopicContinuityBERT*, a BERT model fine-tuned with the *TopicContinuity* dataset. We use a deductive research design with a computational approach, combining quantitative methods for data analysis and model evaluation. Data was collected through web scraping of political speeches from publicly available sources, followed by annotation and feature extraction using rule-based and ML techniques. We fine-tuned a BERT model on a custom dataset to evaluate linguistic features influencing topic continuity and applied explainability tools for qualitative and quantitative analysis. Datasets, models, and code are available in our GitHub repository.

3.1 Datasets

We designed TopicContinuity to be perfectly balanced, with equal representation of both continuity classes. This stratification was key to eliminating bias, incorporating explainability of the linguistic features that characterize topic continuity in political discourse, and attempting the generalizability of our findings. We collected 42K public discourses (in total, we collected 42K speeches, interviews, debates, or similar) using an ad-hoc web-scraping tool (Reyes, 2023a) from American-targeted websites, predominantly from The American Presidency Project (Peters & Woolley, n.d.) and from news websites, government archives, and government agencies' websites. The source websites were of two kinds: (1) publicly accessible, with permission for fair

use purposes and academic research, and (2) proprietary, with granted authorization after request. In both cases, we complied with the terms set by each source to ensure the legal and ethical use of their materials, including appropriate citations and adherence to legal and academic standards.

A comprehensive cleaning procedure on the collected texts was implemented to ensure data quality, including removing URLs, Unicode symbols, speaker labels, bracket annotations, timestamps, and contextual data. We created a *linguistic-rule-based model* (*LRBM*) using spaCy (Honnibal et al., 2020) to extract the following features:

1. *Coreferentiality*: Using spaCy's experimental model for coreference resolution (*en_core_web_trf*), we analyzed coreferential links between two sentences to uncover anaphoric and cataphoric references. We filtered out references that do not span the sentences, focusing only on those contributing to inter-sentence coreferentiality. For example:

> "[1] *The Iraqis have been trying to* ***acquire*** *weapons of mass destruction.* [2] ***That****'s the only explanation for why Saddam Hussein does not want inspectors in from the U.N.*"

The coreferentiality information extracted by spaCy's coreference model from the previous sentence pair shows that the cataphoric reference "*that*" from the second sentence refers to the syntactic root of the first clause in the first sentence, "*acquire*":


```
{
  "coreference": {
    "coreference_group_1": [
      {
        "coref": "acquire",
        "start": 6,
        "end": 7
      },
      {
        "coref": "That",
        "start": 12,
        "end": 13
      }
    ]
  }
}
```


2. *Lexical cohesion*: Using spaCy'slemmanitization and parts of speech (POS), we evaluated if two sentences shared common lexical units that contributed to the thematic unity and flow of discourse, specifically nouns ("NOUN"), proper nouns ("PROPN"), verbs ("VERB"), adjectives ("ADJ"), adverbs ("ADV"), and numerals ("NUM"). We compared the lemmas of important words in both sentences. For example:

> *"[1]* ***African American*** *youth* ***unemployment*** *is the* ***lowest level*** *in the* ***history*** *of our country. [2] And* ***African American unemployment*** *is the* ***lowest level*** *in* ***history****."*

Both sentences share the following lexical units in their lemma form: "*african*", "*american*", "*unemployment*", "*low*", "*level*", and "*history*".

3. *Semantic cohesion*: Leveraging spaCy's semantic similarity feature, we determined whether two sentences shared semantic units at the token level, such as nouns ("NOUN"), proper nouns ("PROPN"), verbs ("VERB"), adjectives ("ADJ"), adverbs ("ADV") and numerals ("NUM"). The process calculated cosine similarity—using the method `.similarity()`—between non-identical tokens to ensure a diverse semantic comparison. Tokens had to exceed a similarity threshold of 0.75 to be considered semantically continuous, ensuring that only tokens with significant semantic relatedness contribute to the continuity between sentences. For example:

> "[1] *And **many** of us grew up in a time when a worker would spend an entire career in the same job, and those days are ending.* [2] *Workers entering the economy today can expect to train and retrain **several** times to keep pace with changed working conditions.*"

Both sentences share the adjectives "*many*" and "*several*" that have the same meaning but use different lexicality.

4. *Syntactic parallelism*: Using spaCy's linguistic features, syntactic parallelism between sentences by exploring the commonality in dependency relationships among individual words, involving an examination of how tokens (words) are syntactically connected to their heads within each sentence, based on their dependency patterns. These token-level patterns are crucial in defining syntactic harmony, which contributes to textual cohesion and parallelism. For example:

> "[1] ***It's** hard to run a business if you're marching to war.* [2] ***It's** not conducive to capital investment.*"

In Figure 1, we see two sentences sharing the same syntactic root, the auxiliary verb "*is*" ("*to be*"), as an indicator of topic continuity. The dependency parsing visualization shows, in both cases, the outgoing arrows coming out from "is", which are the sentences' syntactic roots. This kind of parallelism, focusing on individual token relationships, is frequently used in political discourse, providing a practical application of our findings.

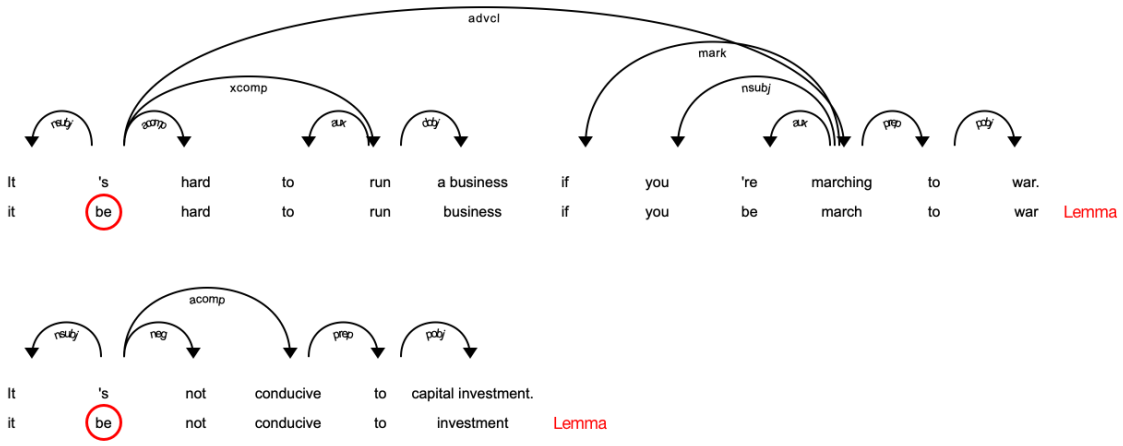

FIGURE 1: Example of Syntactic Parallelism in Two Sentences Using spaCy's Dependency Tree Visualizer.

5. *Transitional cohesion*: We analyzed transitional cohesion using lexicons of transition markers, located as the first token in the second sentence, subdivided into *topic continuity* and *topic shift* markers, as detailed in the Appendix. This systematic categorization allowed us to evaluate

how effectively transitions contribute to the logical progression and coherence of the text. For example:

> "[1] *But we realized the true threats were inside the country, whether it be the Saddamists, some Sunni rejectionists, or Al Qaida that was in there torturing and killing and maiming in order to get their way.* [2] ***And*** *we are making progress when it comes to training the troops*."

We used the LRBM to extract candidate passages. First, the system navigated each political discourse text, sentence by sentence, using a matcher system to find a sentence with at least one political issue of 158 political issues that have been prominent in political discussions and the public sphere in the U.S. over the past 80 years. The matcher system used three different matchers that sought variations of political issues or synonyms, totaling a dictionary of 369 different expressions, implemented in spaCy's custom *named-entity recognition* (NER) component. The three matchers: (1) *Hyphenated term* pattern, which identifies compound words in its lemma and non-hyphenated forms (for example, "*same-sex marriages*" to its lemmatized version "*same sex marriage*"); (2) *Lemmatized* pattern, which allows the system to recognize different forms of a word as the same entity (for example, "*taxes*" and "*tax*"); and (3) *Exact-term* pattern, ensuring precise identification of specific phrases (for example, "*NATO*" and "*N.A.T.O.*"). Then, the matcher checked if the matched political issue played a significant role in the main topic of the sentence., by confirming if their role was a *subject*, *direct object*, *object of a preposition*, *attribute*, or *adverbial clause modifier*. Once a political issue was found in a sentence, the system checked the presence of any of the five topic linguistics features in sentence pairs, upwards first and then downwards, delimiting the range of a passage in the text.

The extraction task resulted in a pool of 8,788 passages, with a minimum of three sentences and a maximum of 10. We randomly selected a split of 800 passages, converted them into sentence pairs, and broke them down into three datasets (train 80%, validation 20%, and test 20%) to fine-tune a prototype model BERT1 as our baseline.

The first annotation round was at the passage level, defining the boundaries of passages about political issues (topics). For that purpose, we created an ad hoc passage annotation tool (Reyes, 2023b) (Figure 2). The tool allowed four actions to annotators over the edited passages: (1) *accept* the passage after modifications, (2) *reject* the passage to flag it as useless, (3) *ignore* the passage to allow another annotator to work on it, and (4) *undo* modifications and start over the passage annotation. This round involved seven annotators and an additional curator to establish the gold standard in case of disagreements.

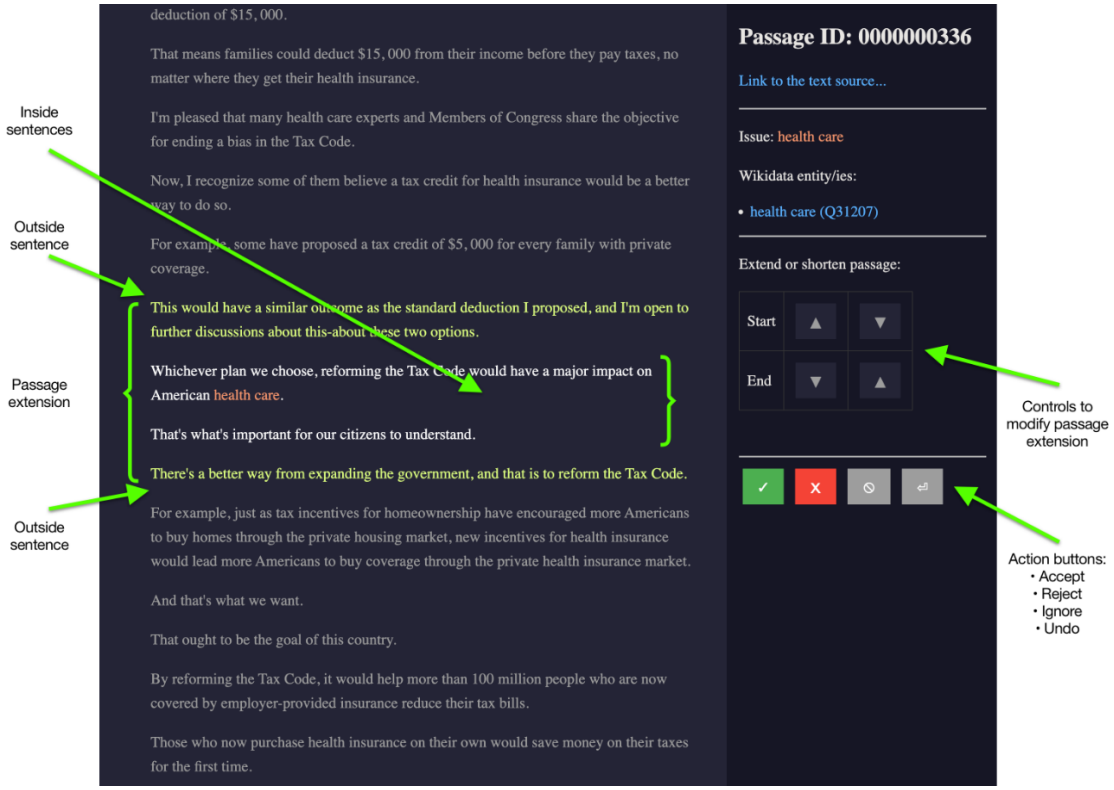

FIGURE 2: UI of the Passage Annotation Tool.

This approach forced annotators to read and understand larger blocks of text, which provided them with a broader context, ensuring that the resulting passages were more likely to be coherent and representative of actual discourse structures. This round ended with 2,881 annotated passages, which we converted into 5,281 sentence pairs (never longer than 512 tokens due to the known token limit of BERT when processing text) by selecting outside sentences (from both the beginning and end of the passages), labeled as the *not continue* class and *inside sentences*, labeled as the *continue* class. For example, from the passage in Figure 2, the following sentence pairs were extracted:

> "[1] *This would have a similar outcome as the standard deduction I proposed, and I'm open to further discussions about this - about this two options*[sic]. [2] *Whichever plan we choose, reforming the Tax Code would have a major impact on American health care.*"

> "[1] *Whichever plan we choose, reforming the Tax Code would have a major impact on American health care.* [2] *That's what's important for our citizens to understand.*"

> "[1] *That's what's important for our citizens to understand.* [2] *There's a better way from expanding the government, and that is to reform the Tax Code.*"

	Text Dataset
Name	*TopicContinuity*
Instances	Sentence Pairs from political discourses
Classes (*)	•Continue (*c*) •Not continue (*nc*)
Number of Instances	2,884 (1,142 *c* / 1,142 *nc*)
Instance Length	Between 8 to 152 tokens
Labels	• "continue" • "not_continue"
Splits/Instances	•Train: 2,306 (79.96%) •Validation: 288 (9.99%) •Test: 290 (10.05%)
Stratification (*)	•Train: 1,153 *c* and 1,153 *nc* •Validation: 144 *c* and 144 *nc* •Test: 145 *c* and 145 *nc*
Metadata	•title (document) •url
Data Period	1939-2023

Note. (*) *c* = continue, *nc*= not continue.

TABLE 1: Datasheet for the TopicContinuity Dataset.

Since the sentence pairs were predominantly from the *continue* class, we allowed a slight imbalance toward that class to fine-tune the prototype model BERT2. In the *second annotation round*, we introduced an anonymous review in the annotation process, where three new annotators were unaware of initial classifications and trained in the five linguistic features, developing documented guidelines and applied examples. This approach demanded a more nuanced linguistics analysis in collaborative annotation sessions that consolidated and extended the guidelines. The IAA analysis with Fleiss's Kappa score of 0.694. Finally, in the *third annotation round*, the same three annotators *pair reviewed* their work, achieving an IAA analysis achieved a Fleiss's Kappa score of 0.812, resulting in the TopicContinuity dataset comprising 2,884 sentence pairs; see datasheet in Table 1. After having a refined understanding of the linguistic features, we defined our ground truth by selecting 290 sentence pairs, 50% for the *continue* class and 50% for the *not continue* class, and fine-tuned TopicContinuityBERT. Annotators were students from an M.Sc.-level seminar voluntarily, with no financial compensation offered. They were fully informed of the study's aims and how their annotations would be used, and they agreed to participate, motivated to gain applied real-world experience and credit in the datasets publication. The annotators' names are credited in the publicly available dataset repository, ensuring proper acknowledgment of their efforts. Our project involved no sensitive data or vulnerable populations, so we did not seek formal institutional approval.

To preserve the linguistic focus of our model—not on thematic and ideological biases—we excluded sentence pairs with strong emotional or ideological tones during the annotation process. Two examples:

> SENTENCE PAIR A: "[1] *America was founded on liberty and independence - not government coercion, domination, and control.* [2] *We are born free, and we will stay free.*"
>
> SENTENCE PAIR B: "[1] *Come to India.* [2] *You will know what racism is.*"

While both sentence pairs stay on the same general topic, their rhetorical tone could mislead the model into thinking that sentiment or ideology—rather than actual linguistic connections—defines topic continuity. By filtering out such examples, we made sure the model learns from structural and wording cues, not from sentiment or controversy. Furthermore, the exclusions of potentially polarizing or sensitive examples contribute to the ethical design of the model, aligning with best practices in NLP and dataset curation (Gebru et al., 2018).

3.2 Experimental Setup

BERT can be fine-tuned for tasks that involve sentence pairs, where these pairs are submitted individually to the model and internally formatted as a single input sequence separated by special tokens (`[SEP]` and `[CLS]`) (Figure 3), which is a method used to maintain context and relational understanding between the two parts of a text (Devlin et al., 2019). When two sentences are fed to BERT, the model sees the–m as a single sequence–`[CLS] sentence A [SEP] sentence B [SEP]`—but can still distinguish which tokens belong to each sentence.

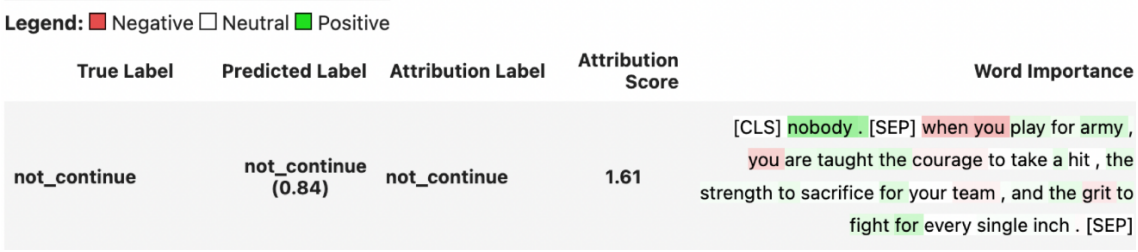

FIGURE 3: Special Tokens [CLS] and [SEP] Added by BERT that Transformers Interpret Leverages to Handle Sentence Pairs Explanations.

The required setup for SMP challenges traditional explainability tools, typically designed to handle tokens, sentences, or text segments independently. However, Transformers Interpret addresses this limitation with *PairwiseSequenceClassificationExplainer*, its explainer specifically designed to interpret the predictions of Transformer models that have been fine-tuned on tasks involving sentence pairs. Transformers Interpret relies on Captum (Kokhlikyan, 2020), an open-source model explainability library built by *Facebook AI Research* specifically for use with *PyTorch* (Paszke et al., 2019), and provides an easier, higher-level interface for explaining *Hugging Face Transformer* models. Using PairwiseSequenceClassificationExplainer, we can examine and identify the contributions of individual tokens in each sentence of the pair towards the model's decision-making process, aiding in understanding TopicContinuityBERT's behavior for sentence-pair classification. During the explainer setup, we had to modify version 0.5.2 of Transformers Interpret because it did not handle the number of tensors of our version of BERT.

We employed the TopicContinuity dataset, divided into training (80%), validation (10%), and testing (10%) subsets, to fine-tune TopicContinuityBERT using sentence pairs separately through the `.encode()`, a method provided by the tokenizer class in the Hugging Face Transformers library on an *Apple Silicon*'s GPU, *Metal Performance Shaders* (*MPS*), utilizing the "*bert-base-uncased*" pre-trained model variant, the *BertForSequenceClassification*, and the PyTorch deep learning framework. We used *Optuna* (Akiba et al., 2019) to find the best model by evaluating maximal performance and minimal overfitting. We monitored the *training and validation losses* closely, employing the early-stop strategy when the training loss ceased to decrease, thereby preventing overfitting (Figure 4). We used the following metrics: *learning rate*, 1.2465928099530177e-05; *batch Size*, 16; *warm-up steps*, 369; *number of epochs*, 3; and *seed*, 42. We used Python's libraries for data manipulation and visualization, such as *Pandas* (The pandas development team, 2020), *Seaborn* (Waskom, 2021), *Matplotlib* (Hunter, 2007), and *Scikit-learn* (Pedregosa et al., 2011).

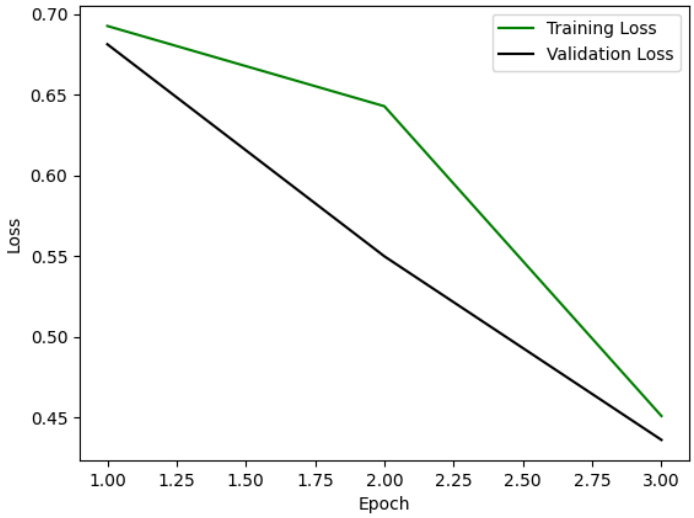

FIGURE 4: Plot of Training and Validation Losses per Epoch During Training of TopicContinuityBERT

For the explainability analysis, we utilized 290 test examples (145 sentence pairs for each class) from the test dataset, previously unseen by the model. As illustrated in Figure 5, we configured a Transformers Interpret explainer connected to TopicContinuityBERT and submitted the sentence pairs for inference. During our observations, we noticed that the tokenizer frequently splits unknown terms into subtokens. To address this issue, we expanded the tokenizer's vocabulary to include these terms as whole tokens, an adjustment that aimed to prevent subtokenization, which we found introduced inconsistency and variability in our token-level analysis, complicating the interpretability of our results.

We combined the LRBM with two rounds of human annotation for the aggregation task. As input, we took each token (and its corresponding score) from the model's predictions and mapped it to one of the five linguistic features based on the token's role in the sentence, irrespective of whether its contribution (polarity) was positive or negative. We employed a cascade ranking strategy: if a token demonstrated coreferentiality, we assigned it to that feature and did not consider it for simpler features further down the list. For instance, if two tokens, "*we*" and "*we*", from consecutive sentences referenced the same entity, they were automatically grouped under coreferentiality—and not, for instance, to lexical cohesion (common lexical units).

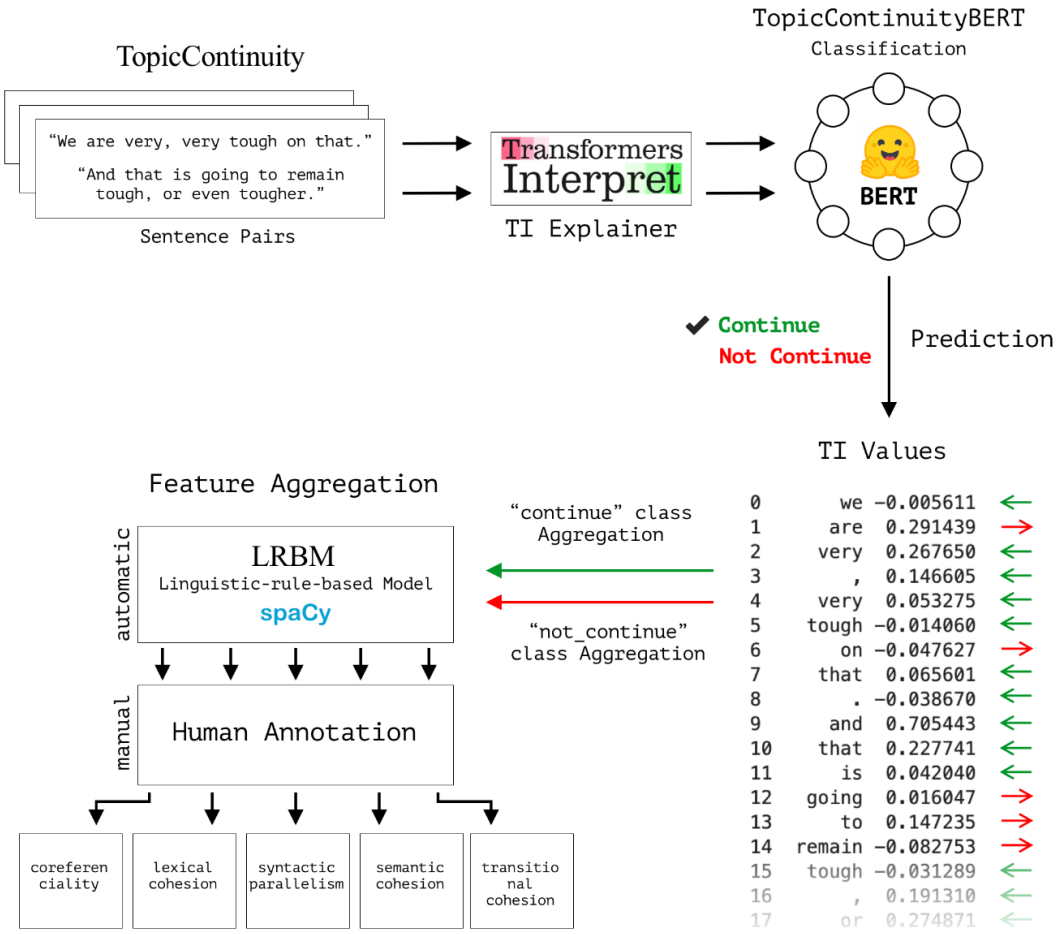

FIGURE 5: Explainability Analysis of TopicContinuityBERT's Behavior using Transformers Interpret.

A set of well-defined guidelines governed this cascade, prioritizing more complex features first, in line with empirical findings that deeper linguistic markers, such as coreferentiality, yield richer insights into topic continuity than simpler ones like lexical repetition (Ledoux et al., 2007). Thus, the ranking of linguistic features was: (1) coreferentiality, (2) syntactic parallelism, and (3) lexical cohesion. Despite being simpler, lexical cohesion still plays a crucial role in textual coherence. We excluded semantic cohesion and transitional cohesion from the cascade because their identification and aggregation were very straightforward. For instance, semantic cohesion always involves multi-token comparisons (as no other feature), while transitional cohesion typically focuses on a single initial token of the second sentence (also as any other feature).

As output, this aggregation yielded two final lists: one for the *continue* class, containing 810 token-value pairs (i.e., tokens mapped to their assigned feature), and one for the *not continue* class, containing 107 token-value pairs. These final lists reflect how TopicContinuityBERT's explanations align with our linguistic feature hierarchy, allowing us to see which tokens (and features) contributed most to the model's classification decisions.

We computed the mean and other descriptive statistics for each linguistic feature separated by class, including all tokens/values—both positive and negative. This approach ensured that our analysis reflected the full spectrum of each token's influence on the model's decision-making

process, capturing both supportive and detractive elements of topic continuity. We plotted overlapping histograms for each feature, with separate curves for the *continue* and *not continue* classes, and the distributions were clearly non-normal. The aggregation process was carried out separately for each class, ensuring that the distributions remained independent and unbiased by class imbalance or overlap. Histograms were plotted using raw frequency counts, which allowed us to observe class-specific value distributions without normalization bias. Finally, we opted for the *Mann–Whitney U* test to compare feature distributions between the classes, given its suitability for detecting location shifts between two independent, non-normally distributed samples.

4. RESULTS AND DISCUSSION

In this section, we present the performance metrics of TopicContinuityBERT and its prototypes, highlighting improvements in accuracy and AUC-ROC through enhanced annotation methods (Section 4.1). We further conduct an explainability analysis using Transformers Interpret, identifying coreferentiality, lexical cohesion, and transitional cohesion as significant contributors to topic continuity, while semantic cohesion and syntactic parallelism are less influential (Section 4.2).

Metric		BERT1			BERT2			TopicContinuityBERT	
Accuracy		0.616			0.852			0.914	
Precision (macro)		0.616			0.852			0.914	
Recall (macro)		0.616			0.852			0.914	
F1 Score (macro)		0.616			0.852			0.914	
AUC-ROC		0.690			0.917			0.960	
		c	*nc*		*c*	*nc*		*c*	*nc*
Confusion Matrix (*)	*c*	190	103	*c*	308	50	*c*	131	14
	nc	120	168	*nc*	56	300	*nc*	11	134
Continue Class									
Precision		0.613			0.846			0.923	
Recall		0.648			0.860			0.903	
F1-score		0.630			0.853			0.913	
Not Continue Class									
Precision		0.620			0.857			0.905	
Recall		0.583			0.843			0.924	
F1-score		0.601			0.850			0.915	

Note. (*) *c* = continue, *nc* = not continue. Across-class metrics are macro and class-wise metrics are not averaged.

TABLE 2: Summary of Performance Metrics of TopicContinuityBERT and Interim Models for Classifying Sentence Pairs into Topic Continue and Not Continue.

4.1 Model Performance

Table 2 shows the performance metrics of TopicContinuityBERT and its two prototypes. BERT1 exhibited modest performance with an accuracy of 0.616, and an AUC-ROC of 0.690. Considering that BERT1 was fine-tuned using sentence pairs extracted automatically using the LRBM, we observed that spaCy's capacities allowed a sophisticated analysis yet were insufficient

for capturing deeper semantic relationships and contextual nuances. BERT2 enhanced these metrics, achieving an accuracy of 0.852 and an improved AUC-ROC of 0.917, meaning that the human intervention using the passage annotation tool played a significant role. TopicContinuityBERT, marks a notable improvement, with its accuracy at 0.914, and a significantly higher AUC-ROC of 0.960

The confusion matrix of BERT1 indicated a relatively balanced distribution of errors with 190 true positives and 103 false negatives for the *continue* class and 120 false positives and 168 true negatives for the *not continue* class. This distribution suggests that while the model could identify instances of both classes, it was equally prone to misclassifying them. The persistently high false positives of BERT2 implied that the model struggled to overpredict the *continue* class despite being more accurate in identifying correct cases. However, TopicContinuityBERT, exhibits a significant reduction in false positives (11) and false negatives (14). The model demonstrated a substantial increase in the accuracy of classifications, with 131 true positives for the *continue* class and 134 true negatives for the *not continue* class.

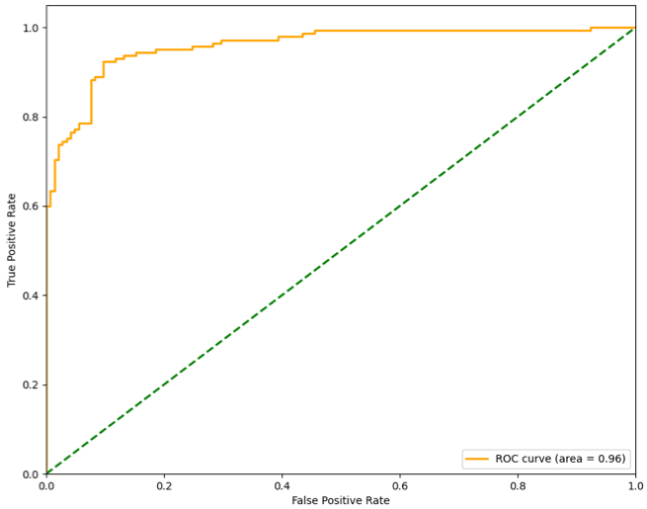

FIGURE 6: Plot of TopicContinuityBERT's ROC Curve

Figure 6 shows an interesting aspect of TopicContinuityBERT: The ROC curve with an AUC of 0.960, indicating that the model has strong discriminative power, with a high true positive rate and a low false positive rate, suggesting its effectiveness in identifying topic continuity.

In Figure 7, the *t-distributed stochastic neighbor embedding* (*t-SNE*) plot of the model's embeddings visually captures the ambiguity inherent in the detection of topic continuity in political discourse, and the overlap between both clusters suggests that the model, while effective, operates in a complex feature space where clear separations are challenging. This overlap could reflect the nuanced and subtle use of language that defines topic continuity, which is not always straightforward or binary. In sum, we can observe the potential and challenges in automated detection of topic continuity that effectively harnesses deep learning to interpret linguistic features, although the task complexity is visible.

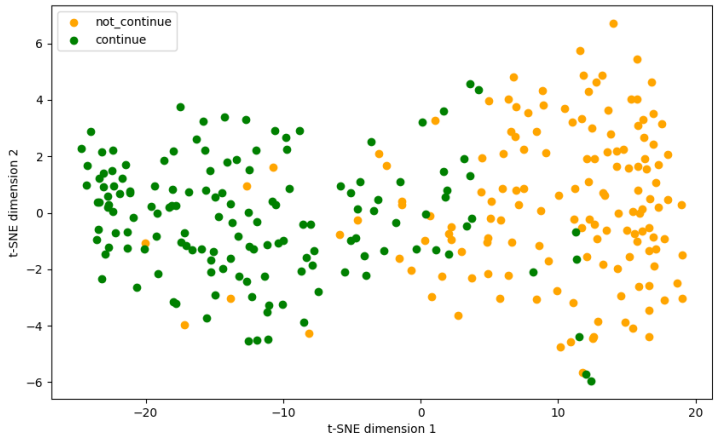

FIGURE 7: t-SNE Plot Embeddings of TopicContinuityBERT

4.2 Explainability Analysis with Transformers Interpret

The descriptive statistics of features computed by Transformers Interpret (Table 3) show that coreferentiality has a mean score significantly higher in the *continue* class (0.160) compared to the not *continue* class (-0.170), suggesting that references linking back or forward toward previously mentioned entities tend to strongly support topic continuity. Similarly, lexical cohesion shows a notably higher mean in the *continue* class (0.129), implying that lexical similarities contribute to perceived continuity, yet with notable variability, suggesting other factors might play a more substantial role in certain contexts.

Feature	Mean	Range	SD	Variance	Skewness	Kurtosis
Continue Class						
Coreferentiality	0.160	1.445	0.216	0.047	0.137	1.648
Lexical cohesion	0.129	1.277	0.189	0.036	0.842	1.696
Semantic cohesion	0.084	1.133	0.158	0.025	0.210	4.454
Syntactic parallelism	0.161	1.451	0.209	0.043	0.355	0.700
Transitional cohesion	0.642	0.941	0.253	0.064	-0.854	-0.381
Not Continue Class						
Coreferentiality	-0.170	1.096	0.207	0.043	0.425	1.967
Lexical cohesion	-0.197	0.929	0.205	0.042	-0.695	0.714
Semantic cohesion	-0.110	1.070	0.271	0.074	-0.515	2.568
Syntactic parallelism	-0.065	0.874	0.258	0.066	-0.290	-0.902
Transitional cohesion	-0.234	0.389	0.181	0.033	0.218	-2.538

Note. Raw aggregated data from TopicContinuity's test dataset is public.

TABLE 3: Descriptive Statistics for the Continue and Not Continue Classes.

Semantic cohesion presents a lower mean score in both classes but remains higher in the *continue* class (0.084 vs. -0.110); however, its high kurtosis in the *continue* class (4.454) suggests that semantic ties, while generally less prominent, can significantly enhance topic continuity when they are present. Syntactic parallelism and transitional cohesion also show clear distinctions between the two classes, particularly with transitional cohesion—which has the highest mean difference (0.642 vs. -0.234). Their presence or absence sharply influences the judgment of continuity. While some features like coreferentiality and transitional cohesion have a more pronounced and straightforward impact, others, like semantic cohesion, contribute more subtly yet remain equally vital.

The histograms in Figure 8 visually confirm these findings, with the *continue* class showing peaks at positive values and the *not continue* class at negative values, especially for transitional cohesion, underscoring its critical role in signaling either the continuation or the segmentation of topics. The histograms also suggest the data deviate from normality, aligning with the noticeable skewness and differences in kurtosis—confirmed by skewness values far from zero and kurtosis values significantly different from 3. These varied distributions across features highlight the complexity of discourse structure, suggesting that effective topic continuity analysis in political texts requires consideration of multiple, interlinked linguistic dimensions.

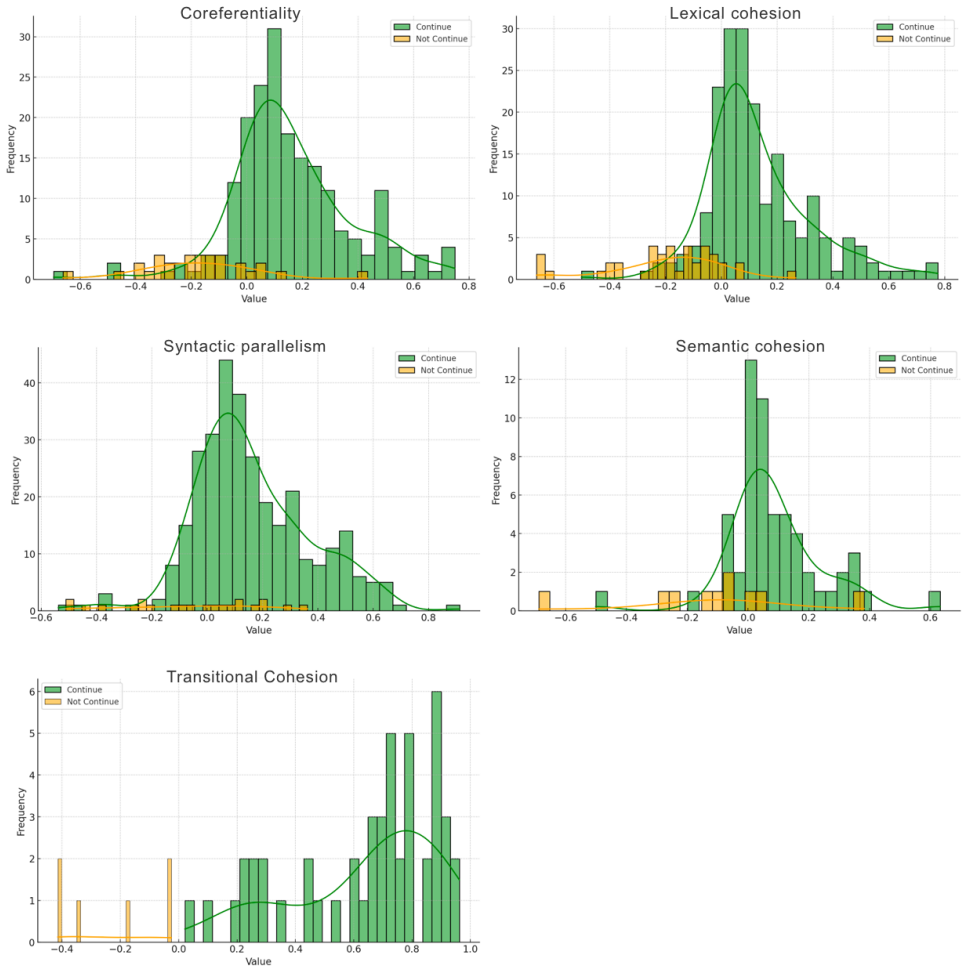

FIGURE 8: Histograms of Data Distribution per Feature Across the Continue and Not Continue Classes.

As seen in Table 4, the results of the Mann-Whitney U test revealed that lexical cohesion, transitional cohesion, and coreferentiality displayed statistically significant differences between the *continue* and *not continue* classes. Specifically, lexical cohesion showed a pronounced difference with a U statistic of 5,483 and a *p*-value of 4×10^{-6}, suggesting its important role in predicting topic continuity. Similarly, transitional cohesion, which plays a pivotal role in connecting ideas, demonstrated a significant difference with a U statistic of 313 and a *p*-value of 8×10^{-6}. Coreferentiality, which involves the use of pronouns and other referential devices to maintain topic continuity, also indicated a significant effect, as evidenced by a U statistic of 2,241 and a *p*-value of 1.1×10^{-5}. Oppositely, semantic and syntactic parallelism did not exhibit a significant difference between both classes, with U statistics of 274 and 2,109, respectively, and *p*-values of 0.788 and 0.619, indicating that —at least within the scope of this dataset—both might not be as influential in predicting topic continuity.

Feature	U Statistic	*p*-Value	Significance
Lexical cohesion	5,483	**< .00001**	**Yes**
Transitional cohesion	313	**< .00001**	**Yes**
Semantic cohesion	274	0.788	No
Syntactic parallelism	2,109	0.619	No
Coreferentiality	2,241	**< .00001**	**Yes**

Note. A *p*-value less than 0.05 is considered statistically significant.

TABLE 4: Summary of the Mann-Whitney U test on Features for Topic Continuity.

In summary, the explainability analysis with Transformers Interpret revealed that coreferentiality, lexical cohesion, and transitional cohesion consistently emerged as strong indicators of topic continuity, echoing their higher positive means and significant Mann–Whitney U test results. In contrast, semantic cohesion and syntactic parallelism, though present, appeared less influential in TopicContinuityBERT. The distribution of values for each feature also deviated from normality, underscoring that discourse-level factors in political texts seldom align with strictly uniform patterns. These findings highlight that while certain linguistic features have an especially pronounced role in signaling topic continuation, effective analysis of political discourse requires an integrative view of multiple, interlinked topic continuity features.

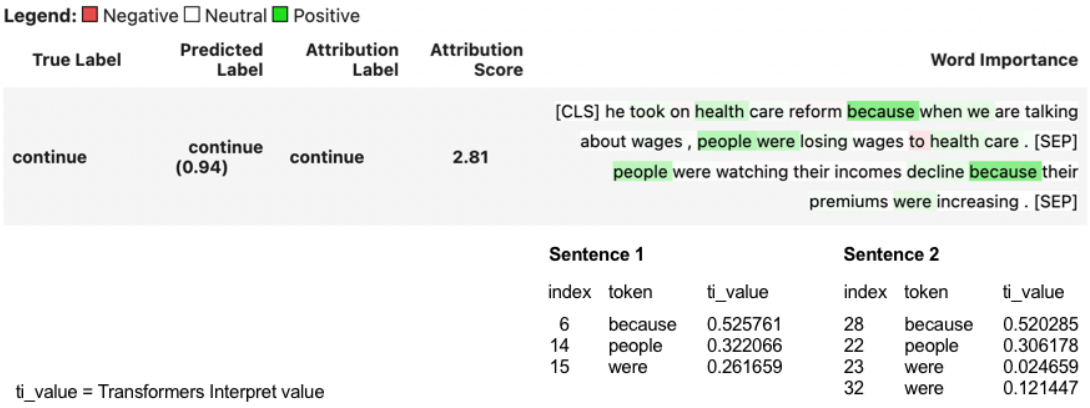

FIGURE 9: Example 1: Explainability Analysis with Transformers Interpret in Topic Continuity Using Sentence Pairs.

ISSN: 2180-1266, https://www.cscjournals.org/journals/IJCL/description.php

Two examples give a more granular glance at TopicContinuityBERT's behavior. In Example 1 (Figure 9), the explainer scores three words as the higher contributors toward the *continue* class: "*because*", "*people*", and "*were*" (duplicated in Sentence 2). The three words, present in seven tokens in both sentences, have a strong role in the prediction, with all having the highest positive values in the sentence pair. This observation confirms the model's reliance on lexical continuity to define topic continuity. Additionally, the token "people" in Sentence 2 is the coreference of "*people*" in Sentence 1, confirming the presence of coreferenciality as a second topic continuity feature.

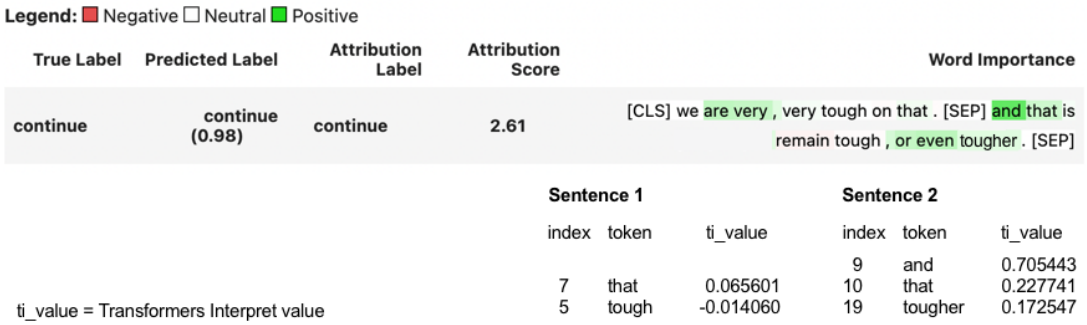

FIGURE 10: *Example 2 part 1: Sentence Pair Before Ablation Analyzed with Transformers Interpret in Topic Continuity.*

Example 2 (Figures 10 and 11) illustrates how TopicContinuityBERT, adapts when a critical word is removed from the input. This ablation exercise serves as a qualitative analysis to uncover how the model shifts its reliance from one feature to another in its output. In Figure 10, we observe the prediction with the original sentence pair, where the token "*and*" in Sentence 2, a coordinating conjunction that links both sentences through the transitional cohesion feature, is scored with the highest value in the prediction of topic continuity.

Figure 11 illustrates the results of the ablation after we removed the token "*and*" from Sentence 2. TopicContinuityBERT adjusted its behavior, now relying on another continuity feature, coreferentiality, scoring the token "*that*" in Sentence 2—which refers to the token "*that*" in Sentence 1—with the highest value in the prediction of topic continuity.

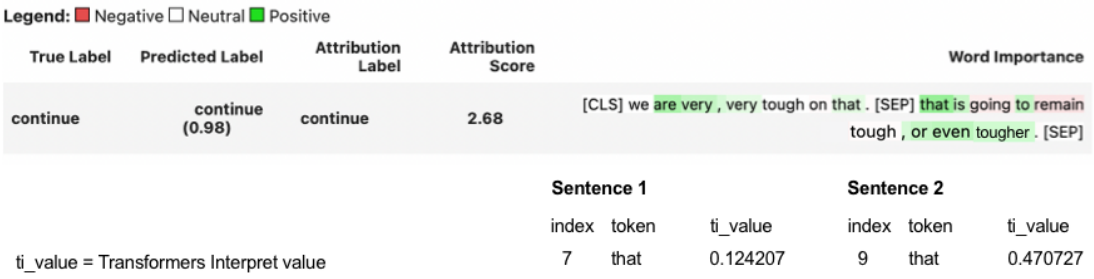

FIGURE 11: Example 2 part 2: Sentence Pair After Ablation Analyzed with Transformers Interpret in Topic Continuity.

To the best of our knowledge, this is the first approach integrating SPM with explainability tools to assess how coreferentiality, lexical cohesion, syntactic parallelism, and transitional cohesion contribute to topic continuity within sentence pairs. Unlike previous studies that focus on broader text segmentation tasks or entailment, our work addresses a clear research gap by targeting the granular analysis of sentence pair coherence. The results demonstrate that TopicContinuityBERT

effectively identifies these linguistic features, offering valuable insights into political discourse analysis. By providing interpretability through explainability techniques, our approach not only improves model transparency but also establishes a foundation for future research in XAI applied to political discourse analysis.

5. CONCLUSIONS

While this study used SPM as a controlled exercise to explore specific linguistic markers of topic continuity, we emphasize that this approach is inherently reductive and not suitable for capturing the full complexity of topic continuity in natural discourse. The simplification into sentence pairs neglects broader discourse context, thematic reintroductions, digressions, and higher-order coherence relations that cannot be fully resolved through pairwise analysis alone.

In our study, while semantic and syntactic units are integral to sentence structure and meaning, we found them contributing less strongly to the continuity of topics in political discussions as the use of lexical cohesion, transitional cohesion, and coreferentiality. Therefore, this analysis answers RQ1 by identifying specific linguistic features critical in predicting topic continuity, offering a valuable setup for further research and model development in political discourse analysis.

Transformers Interpret's capabilities to handle sentence pairs were critical to analyzing quantitatively and qualitatively the role of topic continuity features to predict the presence of topic continuity features. Quantitatively, the tool provided information on the contributions to the model's decision-making process with detailed granularity (tokens, value, and direction), which was critical in the data aggregation for further statistical analysis. Qualitatively, it allowed the analysis of individual token contributions to the model's decision-making process—including suitable visualizations and the convenience of making ablation—enhancing our understanding of how specific words and their contextual use influenced topic continuity predictions. This dual approach verified the model's effectiveness and offered critical insights into the complex interaction of linguistic features in political discourse. This multidimensional analysis shows that Transformers Interpret not only aids in identifying which linguistic features are most crucial for topic continuity but also enhances transparency and interpretability, thereby effectively responding to RQ2 by illustrating how explainability tools can bridge the gap between computational assessments and human-centric interpretations of complex linguistic phenomena.

Establishing the appropriate level of granularity for building a topic continuity dataset using sentence pairs is a complex process: whereas overly strict rules may lead to over-segmentation, too lenient rules could overlook subtle expressions of features critical in the analysis. The inherent ambiguity of political discourse, with its strategic rhetoric and stylistic complexities, further complicates this process, as it often blurs the boundaries between different topic segments. Balancing this granularity was theoretically and practically challenging, especially with observed feature interdependencies. Consequently, the annotation process required highly subjective definitions, which in this study necessitated three rounds of combined automatic and manual annotation, with consistent guidelines and examples developed during collaborative sessions, something visible in the incremental improvement of the Fleiss's Kappa score in the IAA from the baseline.

On the other hand, explainability tools for LLMs, such as Transformers Interpret, are inherently restricted to token-level analysis rather than phrase-level analysis, and while token-level interpretation offers simplicity, it overlooks several features crucial at the phrase level. For instance, (1) in transitional cohesion phrases like "*in conclusion*" or "*on the other hand*"; (2) in coreferentiality references like "*Congress passed a new healthcare bill. This will expand coverage for millions*", where the referred noun is really "*new healthcare bill*", and not simply ("*healthcare*");

(3), or in semantic cohesion, where "*America*" and "*United States*" are semantically the same. Unfortunately, token-level analysis is a standard limitation across all current NLP explainability tools, and although phrasal analysis can be artificially implemented with them, it compromises efficiency and accuracy.
Likewise, the crafted LRBM was an important asset in analyzing automatic topic continuity features in our research; however, we found limitations in how we operationalized the extraction of syntactic parallelism and semantic cohesion features. Our implementation of syntactic parallelism captured only parallelisms of syntactic structures between pairs of tokens (each consisting of a head and dependents) that overlooked more complex syntactic parallelism that defines topic continuity between sentences. For instance, in

> "[1] *African American youth unemployment is the lowest level in the history of our country*. [2] *And African American unemployment is the lowest level in history*."

there are multi-token parallelism impossible to capture by our setup:

- SENTENCE A: [(`subject`) "*African American youth unemployment*" + (`head`) "*is*" + (`complement`) "the lowest level in the history of our country"]
- SENTENCE B: [(`subject`) "*African American unemployment*" + (`head`) "*is*" + (`complement`) "*the lowest level in history*"]

Similarly, our operationalization of the extraction of semantic cohesion used spaCy's semantic similarity feature, which can compare similarity between general terms but is limited to capturing the semantical peculiarities in one specific domain.

Another limitation of our research is the focus on only five features that define topic continuity. Although they are not the only features to evaluate topic continuity, they are important predictors, as evidenced in our quantitative and qualitative (ablation exercise) analysis. Likewise, we acknowledge the limitation of our models to the political sphere within the United States and the American English language and its particular linguistic and cultural context.

The practical implications of our findings include improved explainability of LLMs when applied to complex discourse analysis tasks. These contributions are relevant to NLP researchers focused on explainability, political discourse analysts, and developers seeking to enhance the transparency of AI models applied to text segmentation tasks. Also, our approach opens new avenues for developing explainable AI tools capable of dissecting subtle rhetorical structures in various applications.

Finally, the SPM method used in this study is not only crucial for analyzing linguistic features but also pivotal in enhancing the computational understanding of political discourse. This approach has successfully bridged the gap between computational assessments and human-centric interpretations, offering a powerful framework for future research in topic continuity.

6. APPENDIX

Lexicon of Transitional Cohesion Markers (Leading Words)

1. Topic continuity
   and, so, nor, also, furthermore, moreover, besides, additionally, plus, namely, specifically, first, firstly, secondly, thirdly, subsequently, finally, later, next, afterwards, thereupon, henceforth, because, therefore, thus, hence, indeed, actually, certainly, truly, undoubtedly, clearly, obviously, evidently, naturally, notably, unquestionably, assuredly,

inarguably, decidedly, emphatically, unequivocally, categorically, irrefutably, explicitly, conclusively, essentially

2. Topic shift
   but, or, however, nevertheless, nonetheless, conversely, although, though, despite, instead, whereas, while, yet, contrarily, differently, unlike, contradictorily, still, admittedly, regardless, notwithstanding, albeit, rather, surprisingly, contradictorily, previously, initially, lastly, eventually, until, meanwhile, thereafter, consequently, elsewhere, nearby, opposite, adjacent, beyond, alongside, amid, among, between, across, around, behind, beneath, beside, within, surrounding, over, throughout

7. REFERENCES

Abdalla, M., Vishnubhotla, K., & Mohammad, S. (2023). What makes sentences semantically related? A textual relatedness dataset and empirical study. In A. Vlachos & I. Augenstein (Eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 782–796). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.55.

Abdolahi, M., & Zahedi, M. (2016). An overview on text coherence methods. In *2016 Eighth Conference on Information and Knowledge Technology (IKT)*, 1-5. https://doi.org/10.1109/IKT.2016.7777794.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019*) (pp. 2623-2631). Association for Computing Machinery. https://doi.org/10.1145/3292500.3330701.

AnjaliM, K., & BabuAnto, P. (2014). Ambiguities in Natural Language Processing. *International Journal of Innovative Research in Computer and Comunnication Engineering*, 2, 392-394.

Arase, Y., & Tsujii, J. (2021). Transfer fine-tuning of BERT with phrasal paraphrases. *Computer Speech & Language*, 66, 101164. https://doi.org/10.1016/j.csl.2020.101164.

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London: Routledge.

Batubara, M., Rahila, C., & Ridaini, R. (2021). An Analysis Lexical Cohesion In Jakarta Post News. *Journal of Linguistics, Literature and Language Teaching (JLLLT)*. https://doi.org/10.37249/jlllt.v1i1.278.

Carrell, P. (1982). Cohesion Is Not Coherence. *TESOL Quarterly*, 16(4), 479-488. https://doi.org/10.2307/3586466.

Chomsky, N. (1988). Language and Politics (C. P. Otero, Ed.). Black Rose Books.

Davison, A. (1984). Syntactic markedness and the definition of sentence topic. *Language*, 60(4), 797-846. https://doi.org/10.1353/LAN.1984.0012.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171-4186. https://doi.org/10.18653/v1/N19-1423.

Drew, P., & Heritage, J. (1992). Analyzing talk at work: An introduction. In P. Drew & J. Heritage (Eds.), *Talk at work: Interaction in institutional settings* (pp. 3-65). Cambridge University Press.

Duan, Z., Tan, S., Zhao, S., Wang, Q., Chen, J., & Zhang, Y. (2019). Reviewer assignment based on sentence pair modeling. *Neurocomputing*, 366, 97-108. https://doi.org/10.1016/J.NEUCOM.2019.06.074.

Fan, Y., Jiang, F., Li, P., & Li, H. (2024). Uncovering the potential of ChatGPT for discourse analysis in dialogue: An empirical study. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 16998–17010). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1477/.

Fletcher, C. (1984). Markedness and topic continuity in discourse processing. *Journal of Verbal Learning and Verbal Behavior*, 23(4), 487-493. https://doi.org/10.1016/S0022-5371(84)90309-8.

Gale, W., & Church, K. (1993). A Program for Aligning Sentences in Bilingual Corpora. Comput. Linguistics, 19, 75-102.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv* preprint. https://arxiv.org/abs/1803.09010.

Givón, T. (1983). *Topic continuity in discourse: A quantitative cross-language study*. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.3.

Givón, T. (1995). Coherence in Text vs. Coherence in Mind. In M. A. Gernsbacher & T. Givón (Eds.), *Coherence in Spontaneous Text* (pp. 59-115). John Benjamins Publishing Company. https://doi.org/10.1075/pc.1.2.01giv.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & *Computers*, 36(2), 193–202. https://doi.org/10.3758/BF03195564.

Greenspan, S., & Segal, E. (1984). Reference and comprehension: A topic-comment analysis of sentence-picture verification. *Cognitive Psychology*, 16(4), 556-606. https://doi.org/10.1016/0010-0285(84)90020-3.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Halliday, M. A. K., & Matthiessen, C. (2014). *Halliday's Introduction to Functional Grammar* (4th ed.). Routledge.

Haruno, M., & Yamazaki, T. (1996). High-Performance Bilingual Text Alignment Using Statistical and Dictionary Information. https://doi.org/10.3115/981863.981881.

Honnibal, M., Montani, I., Van Landeghem S., & Boyd A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. https://dx.doi.org/10.5281/zenodo.1212303.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95. https://doi.org/10.1109/MCSE.2007.55.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., &Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for PyTorch. *arXiv*. https://arxiv.org/abs/2009.07896.

Ledoux, K., Gordon, P., Camblin, C., & Swaab, T. (2007). Coreference and lexical repetition: Mechanisms of discourse integration. *Memory & Cognition*, 35, 801-815. https://doi.org/10.3758/BF03193316.

Melamed, I. (1999). Bitext Maps and Alignment via Pattern Recognition. Comput. Linguistics, 25, 107-130.

Newman, E., Stokes, N., Dunnion, J., & Carthy, J. (2005). Textual Entailment Recognition Using a Linguistically-Motivated Decision Tree Classifier. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment (MLCW 2005)*, 372-384. https://doi.org/10.1007/11736790_21.

Orwell, G. (1946). Politics and the English Language. Horizon.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (8026-8037). https://doi.org/10.48550/arXiv.1912.01703.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Louppe, G., Prettenhofer, P., Weiss, R., Weiss, R., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. https://doi.org/10.48550/arXiv.1201.0490.

Peng, Q., Weir, D., & Weeds, J. (2023). Testing Paraphrase Models on Recognising Sentence Pairs at Different Degrees of Semantic Overlap. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, 259-269. https://doi.org/10.18653/v1/2023.starsem-1.24.

Peters, G., & Woolley, J. T. (n.d.). The American Presidency Project. University of California, Santa Barbara. https://www.presidency.ucsb.edu/.

Pi, S.-T., Bagavan, P., Li, Y., Disha, D., & Liu, Q. (2024). Don't shoot the breeze: Topic continuity model using nonlinear naive Bayes with attention. In F. Dernoncourt, D. Preoţiuc-Pietro, & A. Shimorina (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 65–72). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-industry.6.

Pierse, C. D. (2024). *Transformers Interpret*. GitHub repository. https://github.com/cdpierse/transformers-interpret.

Putri, A., & Sudaryat, Y. (2021). Grammatical Cohesion in Moh. Sanoesi's Siti Rayati. *Proceedings of the Fifth International Conference on Language, Literature, Culture, and Education (ICOLLITE 2021)*. https://doi.org/10.2991/assehr.k.211119.007.

Reyes, J. F. (2023a). *webCrawler, a web crawler for political discourse texts* [Source code]. GitHub repository. https://github.com/pacoreyes/webCrawler.

Reyes, J. F. (2023b). *annotationNLP, a web application for annotating NLP datasets* [Source code]. GitHub repository. https://github.com/pacoreyes/annotationNLP.

Ryu, J., & Jeon, M. (2023). An analysis of the inter-grade continuity of the reading passages of high school English mock CSAT tests using Coh-Metrix. *The English Teachers Association in Korea*. https://doi.org/10.35828/etak.2023.29.2.41.

Schiffrin, D. (1994). *Approaches to discourse*. Blackwell Publishers. https://archive.org/details/approachestodisc0000schi.

Shen, G., Yang, Y., & Deng, Z. (2017). Inter-weighted Alignment Network for Sentence Pair Modeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1179-1189. https://doi.org/10.18653/v1/D17-1122.

Sidnell, J. (2010). *Conversation Analysis: An Introduction*. Wiley-Blackwell.

Tang, N., & Moindjie, M. (2024). Lexical Cohesion in English-Chinese Business Translation: Human Translators Versus ChatGPT. *World Journal of English Language*. https://doi.org/10.5430/wjel.v15n2p286.

Tannen, D. (Ed.). (1984). *Coherence in Spoken and Written Discourse*. Ablex Publishing.

The pandas development team (2020). pandas-dev/pandas: Pandas. Zenodo. https://doi.org/10.5281/zenodo.3509134.

Van Dijk, T. A. (1980). *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition (1st ed.)*. Routledge. https://doi.org/10.4324/9780429025532.

Wang, K., Zhao, X., Li, Y., & Peng, W. (2023). M3Seg: A maximum-minimum mutual information paradigm for unsupervised topic segmentation in ASR transcripts. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7928–7934). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.492.

Waskom, M. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, 6(60), 3021. https://doi.org/10.21105/joss.03021.

Xu, S., Shijia, E., & Xiang, Y. (2020). Enhanced attentive convolutional neural networks for sentence pair modeling. *Expert Systems with Applications*, 151. https://doi.org/10.1016/j.eswa.2020.113384.

Yang, Y., Qi, S., Liu, C., Wang, Q., Gao, C., & Xu, Z. (2023). Once is enough: A light-weight cross-attention for fast sentence pair modeling. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2800–2806). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.168.

Yin, W., Schütze, H., Xiang, B., & Zhou, B. (2016). ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs. In *Transactions of the Association for Computational Linguistics*, 4, 259-272. https://doi.org/10.1162/tacl_a_00097.

Yu, S., Su, J., & Luo, D. (2019). Improving BERT-Based Text Classification With Auxiliary Sentence and Domain Knowledge. *IEEE Access*, 7, 176600-176612. https://doi.org/10.1109/ACCESS.2019.2953990.

Yu, R., Lu, W., Lu, H., Wang, S., Li, F., Zhang, X., & Yu, J. (2021). Sentence pair modeling based on semantic feature map for human interaction with IoT devices. *International Journal of Machine Learning and Cybernetics*, 12, 3081-3099. https://doi.org/10.1007/s13042-021-01349-x.

Zhao, J., Lan, M., Niu, Z., & Lu, Y. (2015). Integrating word embeddings and traditional NLP features to measure textual entailment and semantic relatedness of sentence pairs. In *2015 International Joint Conference on Neural Networks (IJCNN)*, 1-7. https://doi.org/10.1109/IJCNN.2015.7280462.