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

Institute of Computer Science, Brandenburgische Technische Universität Cottbus-Senftenberg

Juan-Francisco Reyes

pacoreyes@protonmail.com

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

Topic continuity significantly influence the structure and interpretation of conversations, especially within the complex field of political communication (Givón et al., 1983; Fletcher, 1984; AnjaliM & BabuAnto, 2014). What defines these continuity is crucial for understanding political narratives and their impact on public discourse. This study defines *topic continuity* as the presence of specific linguistic markers that suggest a sustained subject or theme between two consecutive sentences within political discourse. 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.

The inherent ambiguity of political language, characterized by its strategic rhetoric and stylistic complexities, 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 BERT (Bidirectional Encoder Representations from Transformers)'s capacity to handle sentence pairs for sequence classification and the Transformers Interpret library's (Pierse, 2024) capacity to explain the features that semantically 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. This approach allows us to set a baseline for further development and 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, 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 several pivotal research questions:

- *RQ1*: Which linguistic features contribute to predict 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.

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. These research niches collectively explore aspects of cohesion and coherence, which are vital for maintaining a seamless flow of topics within a text, ensuring that the information presented is logically and semantically interconnected (Carrell, 1982; Abdolahi & Zahedi, 2016).

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 (Tannen, 1984; Drew & Heritage, 1992; Schiffrin, 1994; Sidnell, 2010), 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—being against something—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. Davison (1984) used sentence pairs to analyze topic continuity, exploring the relationship between linguistic features of sentence topics and their role in discourse using syntactic and semantic properties, using sentence pairs to analyze topic continuity. Likewise, Greenspan & 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 (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 (Yin et al., 2016; Peng et al., 2023). 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.

By traditional linguistic features in topic continuity, we mean the markers that define the continuity of a topic; 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 continuity. Van Dijk (1980) explored semantic cohesion, suggesting that changes in the semantic field or theme from one sentence to another can mark topic continuity, such as shifting from discussing a historical event to detailing a personal anecdote. Finally, Halliday and Matthiessen (2014) discussed transitional cohesion, where the use of conjunctions and transitional phrases (e.g., "*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.

3. Models

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.

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 public discourses using an ad-hoc web-scraping tool from American-targeted websites, predominantly from The American Presidency Project (Peters & Woolley, n.d.) and from news websites, government archives, and government agencies' websites. From the collected 42K speeches, interviews, debates, or similar. 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 to extract the following features:

1. *Coreferentiality*: Using spaCy's experimental model for coreference resolution (*en_coreference_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's lemmanitization 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.

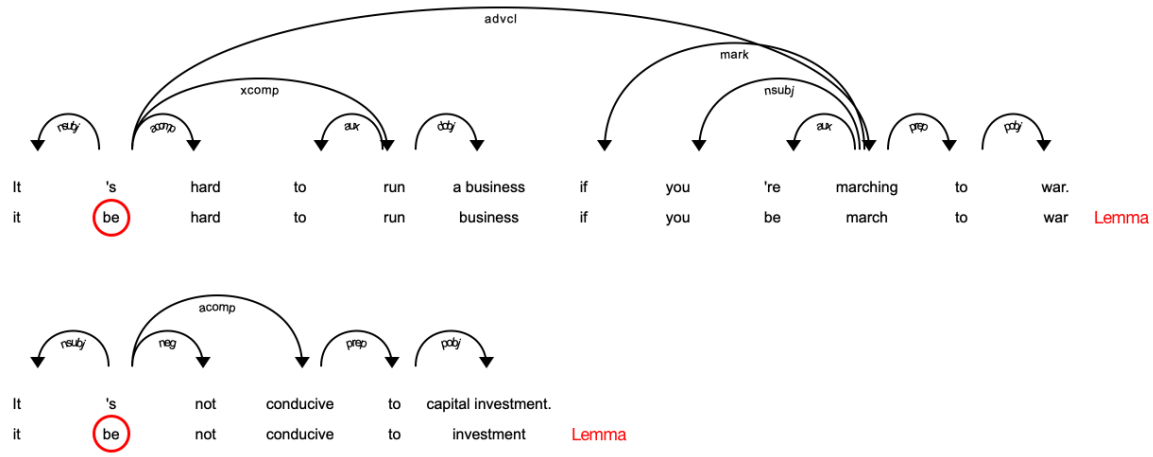
4. *Syntactic parallelism*: Using spaCy's linguistic features, syntactic parallelism between sentences by exploring the commonality in dependency relationships among individual words, involving an examination of how tokens (words) are syntactically connected to their heads within each sentence, based 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*."

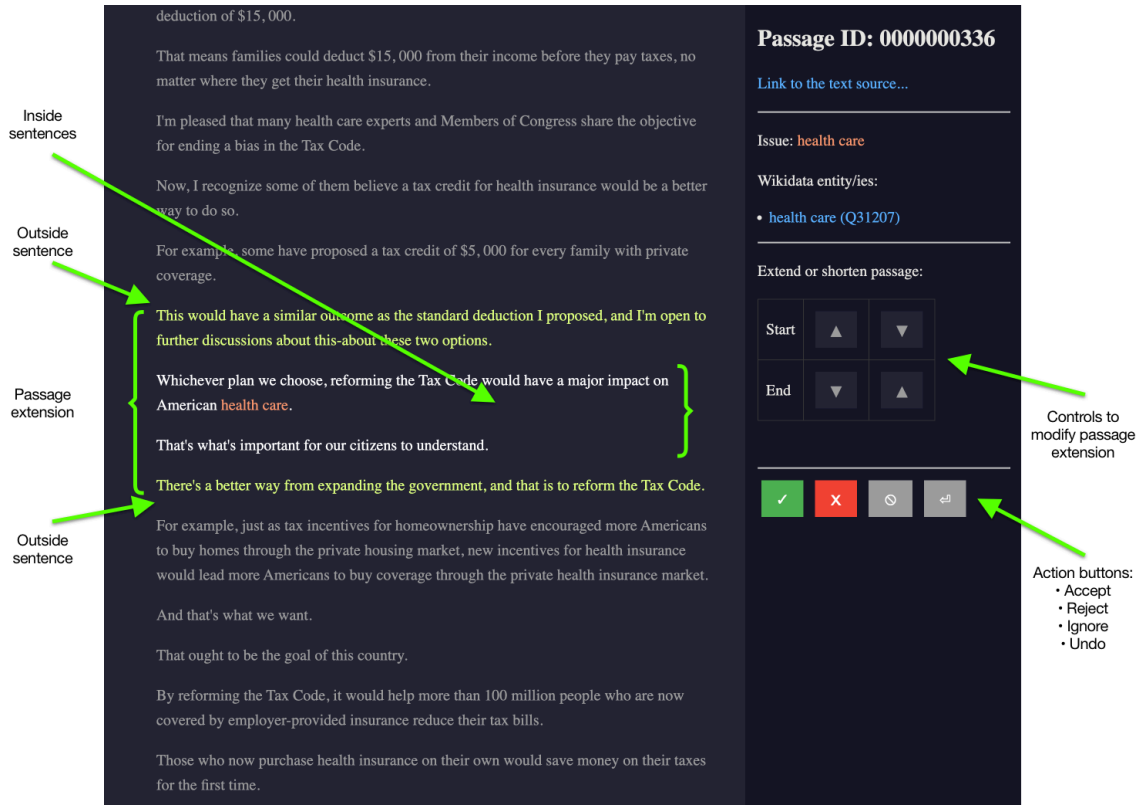
We used the LRBM to extract candidate passages. First, the system navigated each political discourse text, sentence by sentence, using a matcher system to find a sentence with at least one political issue of 158 political issues that have been prominent in political discussions and the public sphere in the U.S. over the past 80 years. The matcher system used three different matchers that sought for variations of political issue 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*, the annotation 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 (Figure 2) that allowed annotators to improve the passages' extensions by extending or shortening them. 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 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) 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*. [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*."

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 a blind 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 achieved a Cohen's Kappa score of 0.724. Finally, in the *third annotation round*, the same three annotators *pair reviewed* their work, achieving an IAA analysis achieved a Cohen's Kappa score of 0.837, 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.

Table 1

Datasheet for the TopicContinuity Dataset.

	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* ad 144 *nc* • Test: 145 *c* and 145 *nc*
Metadata	• title (document) • url
Data Period	1939-2023

Note. (*) *c* = continue, *nc* = not continue. The dataset is freely available in Hugging Face: (1) TopicContinuity, DOI: 10.57967/hf/2756 and (2) GitHub, https://github.com/pacoreyes/topic-continuity.

During the annotation process, we excluded examples that established negative or positive biases toward a concept, for instance, sentence pairs like:

> "[1] *America was founded on liberty and independence - not government coercion, domination, and control*. [2] *We are born free, and we will stay free*."

> "[1] *Come to India*. [2] *You will know what racism is*."

3.2 Experimental Setup

Models like BERT can be fine-tuned for tasks that involve sentence pairs, where these pairs are submitted separated 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 (Devlin et al., 2019). This setup challenges traditional explainability tools, typically designed to handle tokens/sentences/texts independently. 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. With 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 BERT models.

Figure 3

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

Legend: Negative Neutral Positive

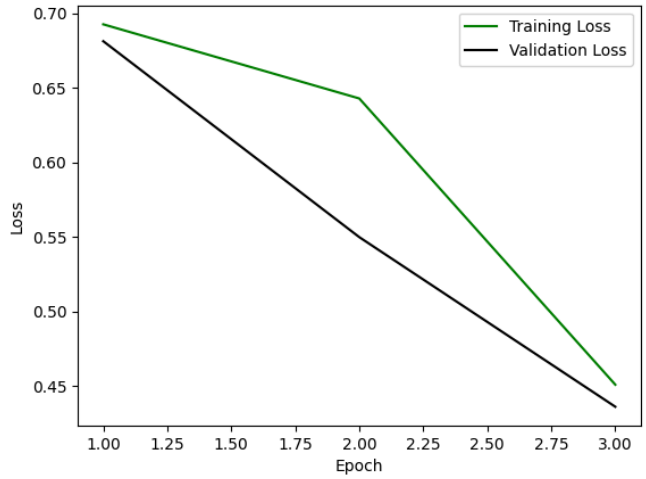
True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
not_continue	not_continue (0.84)	not_continue	1.61	[CLS] nobody . [SEP] when you play for army , you are taught the courage to take a hit , the strength to sacrifice for your team , and the grit to fight for every single inch . [SEP]

We employed the TopicContinuity dataset, divided into training (80%), validation (10%), and testing (10%) subsets, to fine-tune TopicContinuityBERT using sentence pairs separately using 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 (Paszke et al., 2019). 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*, 4; 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 split 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.

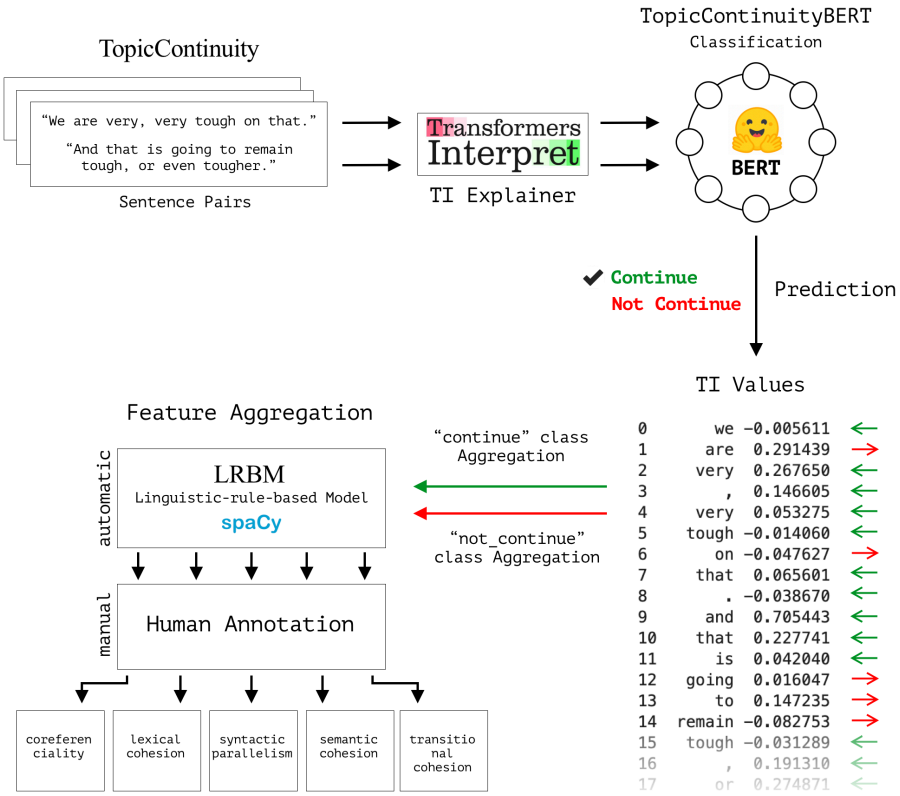
We combined the use of the LRBM with two rounds of human annotation for the aggregation task, incorporating tokens (and their scores) into features based on the token's role in the sentence's structure, regardless of score polarity. For example, if a coreference involved two tokens from different sentences, both tokens and their corresponding values—positive for the continue class and negative for the not continue class—were aggregated under the feature coreferentiality. The human aggregation process followed well-defined guidelines prioritizing linguistic features based on their complexity, asserting that features revealing deeper semantic and syntactic relationships hold greater interpretive value. These guidelines were established based on empirical evidence suggesting that more complex interactions, such as coreferenciality, provide more significant insights into sentence continuity than simpler lexical repetitions (Ledoux et al., 2007). For instance, if the tokens "*we*" and "*we*" were present in both sentences referring one to another, they were aggregated to coreferentiality and not to lexical cohesion.

The prioritization of linguistic features ranked as follows: (1) coreferentiality, (2) syntactic parallelism, and (3) lexical cohesion, which, while simpler, still contributes to the overall textual coherence. Since semantic cohesion involves interactions between two different tokens, and transitional cohesion focuses on a single token in the second sentence, they were not included in the ranking. This process resulted in two lists: the continue class with 810

tokens/values predicted by TopicContinuityBERT as continue, and the not continue class with 107 tokens/values predicted as not continue.

Figure 5

Explanability Analysis of TopicContinuityBERT's Behavior using Transformers Interpret



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 two overlapping histograms representing both classes, analyzed the results; and we did not find data normally distributed, and the aggregation process ensured the data of both classes were independent of each other; we opted for the non-parametric test Mann-Whitney U to compare the medians between both classes to determine if one of them tends to have higher values.

4. Results and Discussion

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 trained using sentence pairs extracted automatically using the LRBM, we observed that spaCy's capacities allow 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.

Table 2

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

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, *o* = not continue. (1) Across-class metrics are macro and class-wise metrics are not averaged. (2) TopicContinuityBERT, DOI: 10.57967/hf/2757, is freely available in Hugging Face.

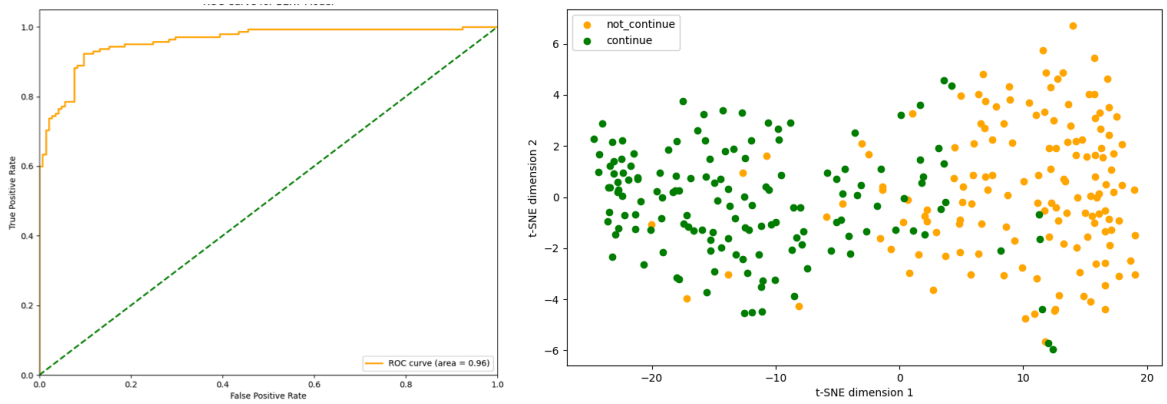
The confusion matrix of BERT1 indicated a relatively balanced distribution of errors with 190 true positives and 103 false negatives for the continue class and 120 false positives and 168 true negatives for the not continue class. This distribution suggests that while the model could identify instances of both classes, it was equally prone to misclassifying them. The persistently high false positives of BERT2 implied that despite being more accurate in identifying correct cases, the model struggled to overpredict the continue class. 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 shows two aspects of TopicContinuityBERT: (1) 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; and, (2) the 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 6

TocpicContinuityBERT's ROC Curve and t-SNE Plot Embeddings



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 support the continuity of the topic strongly. Similarly, lexical cohesion shows a slightly 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.

Table 3

Descriptive Statistics for the Continue and Not Continue Classes.

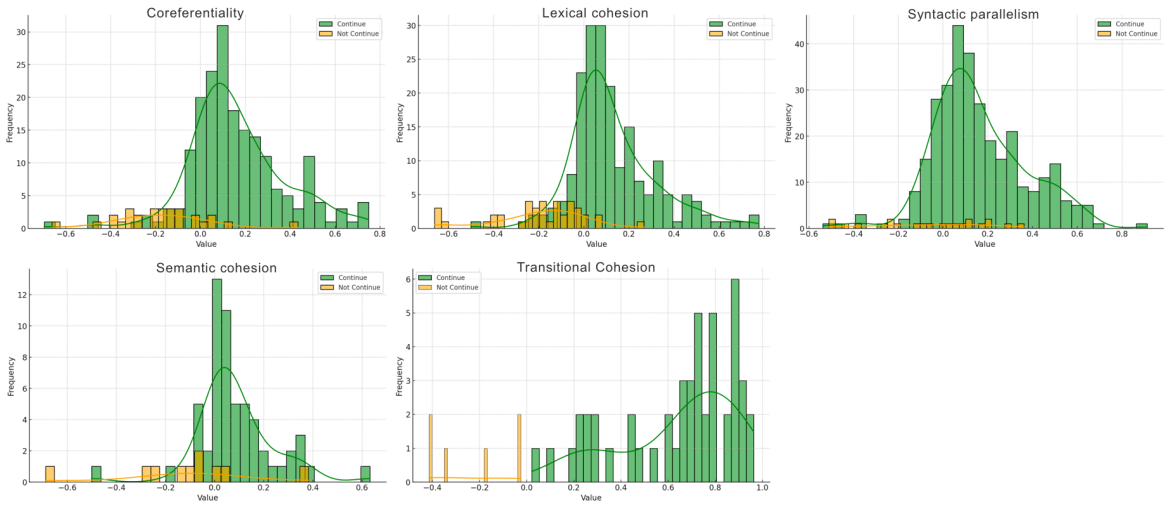
Feature	Mean	Range	SD	Variance	Skewness	Kurtosis
Continue Class						
Coreferentiality	**0.160**	1.445	0.216	0.047	0.137	1.648
Lexical cohesion	**0.129**	1.277	0.189	0.036	0.842	1.696
Semantic cohesion	**0.084**	1.133	0.158	0.025	0.210	**4.454**
Syntactic parallelism	0.161	1.451	0.209	0.043	0.355	0.700
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.

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 equally vitally.

Figure 7

Histograms of Data Distribution per Feature Across the Continue and Not Continue Classes



The histograms in Figure 7 visually confirm these findings, with the continue class showing peaks at positive values and the not continue class at negative values, which is clearly pronounced in transitional cohesion, indicating their critical role in signaling either the continuation or the segmentation of topics. The histograms suggest the data deviate from normality, as they are not perfectly bell-shaped, which aligns with the noticeable skewness and differences in kurtosis, something confirmed with the skewness values, far from zero in all features, and kurtosis values significantly different from 3. The varied skewness and kurtosis across features highlight the complexity of the discourse structure, suggesting that effective topic continuity analysis in political texts requires consideration of multiple, interlinked linguistic dimensions.

Table 4

Summary of the Mann-Whitney U test on Features for 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

Feature	U Statistic	*p*-Value	Significance
Coreferentiality	2,241	**< .00001**	**Yes**

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

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 of 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.

Two examples give a more granular glance at TopicContinuityBERT's behavior. In Example 1 (Figure 8), 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 8

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

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
continue	continue (0.94)	continue	2.81	[CLS] he took on health care reform because when we are talking about wages , people were losing wages to health care . [SEP] people were watching their incomes decline because their premiums were increasing . [SEP]

Sentence 1

index	token	ti_value
6	because	0.525761
14	people	0.322066
15	were	0.261659

Sentence 2

index	token	ti_value
28	because	0.520285
22	people	0.306178
23	were	0.024659
32	were	0.121447

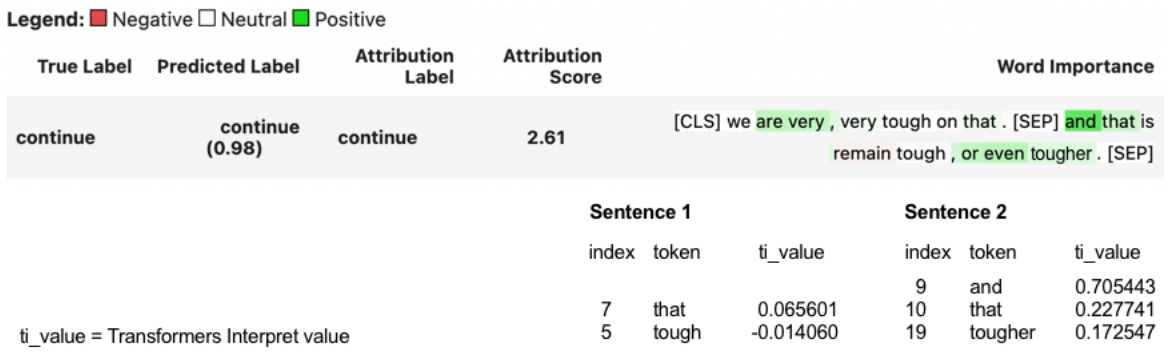
ti_value = Transformers Interpret value

Example 2 (Figures 9 and 10) 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 9, 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 9

Example 2 part 1: Sentence Pair Before Ablation Analized with Transformers Interpret in Topic Continuity.



Legend: Negative Neutral Positive

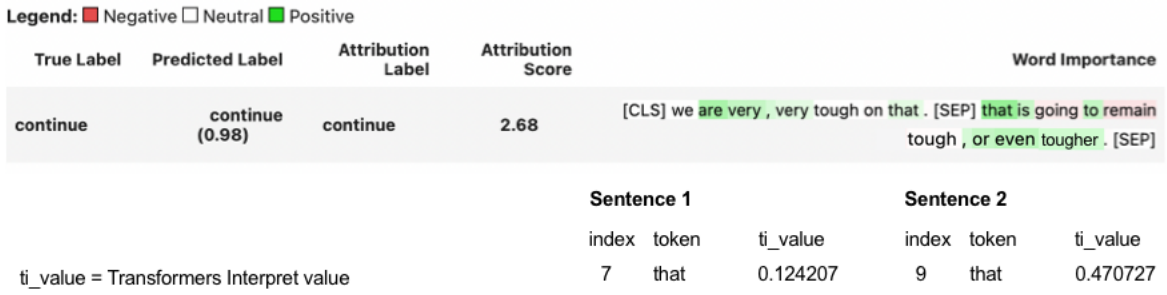
True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
continue	continue (0.98)	continue	2.61	[CLS] we are very , very tough on that . [SEP] and that is remain tough , or even tougher . [SEP]

	Sentence 1			Sentence 2	
	index	token	ti_value	index	token	ti_value
				9	and	0.705443
	7	that	0.065601	10	that	0.227741
ti_value = Transformers Interpret value	5	tough	-0.014060	19	tougher	0.172547

Figure 10 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 10

Example 2 part 2: Sentence Pair After Ablation Analized with Transformers Interpret in Topic Continuity.



Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
continue	continue (0.98)	continue	2.68	[CLS] we are very , very tough on that . [SEP] that is going to remain tough , or even tougher . [SEP]

	Sentence 1			Sentence 2	
	index	token	ti_value	index	token	ti_value
ti_value = Transformers Interpret value	7	that	0.124207	9	that	0.470727

5. Conclusions

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. 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 Cohen's Kappa score in the IAA.

Explainability tools for NLP models, such as Transformers Interpret, are inherently restricted to token-level analysis rather than phrase-level analysis. While token-level interpretation offers simplicity, it overlooks several features crucial at the phrase level. These include

transitional cohesion phrases like "*in conclusion*" or "*on the other hand*", coreferentiality as seen in examples like "*Congress passed a new healthcare bill. This will expand coverage for millions*", and semantic cohesion, exemplified by the contextual similarity between "*America*" and "*United States*" in discussions about the American public sphere. 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.

The LRBM that we developed was an important asset in analyzing automatic topic continuity features; 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 its dependent) that overlooked more complex syntactic parallelism that defines topic continuity between sentences. 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 the domain of American politics. The manual annotation followed the same patterns due to the simplicity of the analysis, but a more nuanced analysis is possible, demanding a higher cognitive load, which should be considered.

We also acknowledge another limitation of our research: the focus on only five features that define topic continuity in political discourse. Although they are not the only features to evaluate topic continuity, we consider them important predictors, as evidenced in our quantitative and qualitative (ablation exercise) analysis. 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, 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.

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.

Appendix

Referential Pronouns

1. Demonstrative pronouns

   this, *these*, *those*, *that*.

2. Personal pronouns

   i, *me*, *my*, *mine*, *we*, *us*, *our*, *ours*, *you*, *your*, *yours*, *he*, *him*, *his*, *she*, *her*, *hers*, *it*, *its*, *they*, *them*, *their*, *theirs*

3. Reflexive pronouns

   myself, *yourself*, *himself*, *herself*, *itself*, *ourselves*, *yourselves*, *themselves*

4. Relative pronouns

   who, *whom*, *whose*, *which*

5. interrogative pronouns

   what, *which*, *who*, *whom*, *whose*

6. Indefinite pronouns

   anyone, *anything*, *everybody*, *everything*, *someone*, *something*, *none*, *nothing*

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*

References

1. 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.

2. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. In *KDD*. https://doi.org/10.48550/arXiv.1907.10902.
3. AnjaliM, K., & BabuAnto, P. (2014). Ambiguities in Natural Language Processing. *International Journal of Innovative Research in Computer and Comunnication Engineering*, 2, 392-394.
4. 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.
5. Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London: Routledge.
6. Carrell, P. (1982). Cohesion Is Not Coherence. *TESOL Quarterly*, 16(4), 479-488. https://doi.org/10.2307/3586466.
7. Davison, A. (1984). Syntactic markedness and the definition of sentence topic. *Language*, 60(4), 797-846. https://doi.org/10.1353/LAN.1984.0012.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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. http://dx.doi.org/10.1075/tsl.31.04giv.
13. 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.
14. Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
15. Halliday, M. A. K., & Matthiessen, C. (2014). *Halliday's Introduction to Functional Grammar* (4th ed.). Routledge.
16. 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.
17. 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.

18. 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.

19. 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)*.

20. 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.

21. 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.

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

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

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

25. 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.

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

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

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

29. 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.

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

31. 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.

32. 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.

33. 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.
34. 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 loT devices. *International Journal of Machine Learning and Cybernetics*, 12, 3081-3099. https://doi.org/10.1007/s13042-021-01349-x.
35. 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.