Juan-Francisco Reyes

AI Explainability in Classifying Political Speeches and Interviews

Abstract

This study applies explainable AI techniques to understand the linguistic features involved in classifying speeches and interviews in political discourse, a field where transparency is sensitive. Using a *feature-based* Linguistic-Rule-Based Model (*LRBM*), *logistic regression*, *Transformer-based models*, and *SHAP values*, we create a more interpretable version of the predictions made by BERT models in this *natural language processing* (*NLP*) binary classification task. The study explores the role that recognizable linguistic features play in both feature-based and neural models. Specifically, it examines the extent to which BERT models depend on linguistic structures for their predictions, using *NER anonymization* to reduce reliance on thematic context. Built on findings from classic and modern linguistic literature, and in addition to improving the interpretability of neural models, the study highlights the identification of important *global* "*political discourse features*" that distinguish speeches and interviews: nominalization frequency, discourse marker frequency, personal pronoun frequency, and interjection frequency.

1. Introduction

This study proposes a binary classification model that distinguishes speeches from interviews, using a *Linguistic-Rule-Based Model* (*LRBM*) and SHAP (*Shapley Additive Explanations*) (Lundberg & Lee, 2017) to explain the Transformer-based model's behavior. Speeches are unidirectional, allowing a speaker to address an audience without direct interruption (*monologic*), whereas interviews are bidirectional, marked by an interactive exchange between interviewer and interviewee (*dialogic*). By examining these distinct discourse types, our research aims to uncover their significant linguistic-structural differences and explore their impact on the explainability of classification tasks in political discourse using *BERT* (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019) models.

Feature-based models, while providing a clear and transparent framework, often fail to deal with the complexities of natural language. While *deep learning* models like BERT have succeeded substantially in text classification, their "*black box*" nature presents explainability challenges (Akpatsa et al., 2021; Castelvecchi, 2016; Lei et al., 2021; Raskin & Harris, 2023), especially vital in political discourse analysis (Chilton, 2004; Zafar et al., 2021a).

One motivation of our research is the precise classification between speeches and interviews, which has profound implications for several subsequent computational linguistic tasks in political discourse analysis. For instance, it can significantly enhance speaker attribution, where the authorship of a passage or statement should be attributed to a specific politician. In that case, the correct discrimination, whether a text has more than one participant, is crucial. Other potential uses of this classification extend from *stylometry* analysis of monologues to determining speech acts and

turn-taking in dialogues. This significant challenge is, hence, far from trivial. Misclassifications or inaccuracies can distort the understanding of a politician's discourse, leading to incorrect interpretations or conclusions. We approach this task with an orientation towards creating a cohesive and efficient analytical pipeline, allowing for integration with other downstream NLP tasks, such as information extraction, information retrieval, text mining, or other text classification tasks.

For a more structured approach to this problem, we studied the taxonomy proposed by Mikhail Bakhtin (1981) and widely used in linguistic and literary studies: *monologism*, monologic discourse (speeches and similar), and *dialogism*, dialogic discourse (interviews and similar). We analyze ten common linguistic features, including *sentence length*, *word length*, *sentence complexity*, *personal pronoun frequency*, *passive voice frequency*, *lexical word frequency*, *nominalization frequency, interjection frequency*, *modal verb frequency*, and *discourse marker frequency*. These features are operationalized computationally using *spaCy* (Honnibal et al., 2020) and four lexicons (see Appendix A) as part of an LRBM that offers high precision and explainability. After measuring features of both discourse classes using the LRBM, we employ a logistic regression (LR) model to define baseline and feature importance and their impact on discourse classification, providing a foundation for examining how the BERT model leverages these features.

Once the importance of features is established, the BERT model, *fine-tuned* with the text datasets, is scrutinized to determine the extent to which it relies on the selected linguistic features for classification, comparing results between a BERT model fine-tuned with a NER anonymized version of the dataset and another with the original dataset, with more contextual information. This approach is intended to compel the model to prioritize linguistic structures over contextual data, addressing concerns that BERT might otherwise rely on recurrent themes rather than linguistic patterns. Utilizing *SHAP* analysis, this study aims to elucidate the weight or influence of specific linguistic features on the neural model, aligning their importance with the explainability of its decisions.

The research is driven by two critical questions:

- *RQ1*: In the task of classifying a political text as a speech or an interview, how can the BERT model's decisions be explained—particularly with respect to which linguistic features most strongly influence the model's predictions?

- *RQ2*: How does fine-tuning a BERT model with anonymized data influence its reliance on linguistic structures over thematic context in classifying political speeches and interviews compared to a model fine-tuned on non-anonymized data?

The contribution of this work is three-fold: (1) our study is the first to deepen the classification of speeches and interviews by leveraging interdisciplinary research combining computational techniques with political science and linguistics; (2) we advance the field of *XAI* (*Explainable Artificial Intelligence*) by presenting an innovative framework that clarifies the decision-making processes of LLMs in the nuanced domain of political discourse and the specific task of text genre classification; and (3) our study contributes with two fine-tuned (BERT) models and two datasets that may help to develop more accurate political analysis tools, enriching linguistic research using state-of-the-art language technology.

The remainder of this paper is structured as follows: Section 2 (Related Work) reviews linguistic and AI-based classification approaches. Section 3 (Models) describes dataset construction, linguistic feature extraction, and the models used (LRBM, Logistic Regression, BERT). Section 4 (Results) presents the results, including linguistic distinctions, model performance, and SHAP-based explainability analysis. Section 5 (Discussion) discusses the findings and their implications, including model behavior under anonymization. Finally, Section 6 (Conclusions) concludes the study and outlines limitations and potential future research directions.

2. Related Work

Due to the scarce specific literature on the linguistic structures of speeches and interviews, we extended our review to the wider categories of modes of communication, monologic and dialogic, using it as a proxy to understand and operationalize the linguistic features of speeches and interviews.

In one of the first studies establishing that speakers adopt different discourse strategies in monological (one speaker) and dialogical (two or more speakers) settings, Gumperz (1982) analyzed language use across social and cultural contexts. Subsequent studies have examined the linguistic features of discourse texts concerning the number of speakers (Kashiha, 2021; Koplenig, 2019; Mauranen, 2023; Mendhakar, 2022; Wells, 2006; Zare & Tavakoli, 2016). Biber et al. (1999) and Hirst (2001) reported that monologic discourse, like speeches, often uses a rich and varied vocabulary, fewer personal pronouns, more complex sentences, and more passive voice, all to convey detailed information accurately. On the other hand, they reported that dialogic discourse, like interviews, typically features simpler language, more personal pronouns, straightforward verb forms and tenses, lower lexical density, and shorter sentence length to make back-and-forth communication easier.

Sentence length and word length are closely related and play a crucial role in the classification of texts. This relationship can be leveraged to distinguish between different modes of discourse, such as monologic and dialogic (Grzybek et al., 2006). McCarthy & Carter (1995) pointed out that conversational language usually employs shorter words. Biber & Finegan (1994) elucidated that dialogic discourse favors shorter sentences, aligning with the conversational, interactive, and real-time processing demands of spoken interactions. Larsson & Kaatari (2020) found a clear distinction in word length between monologic and dialogic texts, asserting that academic and formal texts (often monologic) utilize longer, more complex words to convey in-depth information and maintain a formal tone. Biber (1992) claims that complex sentences are often employed in individual speeches, tailoring to detailed, explorative, and descriptive communication.

Likewise, studies by Biber (1988) into spoken grammar revealed that dialogic discourse tends to have a lower lexical density, potentially arising from its interactive, immediate, and clear communication needs. Conversely, Amelia et al. (2020) found higher lexical density and grammatical intricacy in speeches due to the considerable amount of information using many lexical items as the proportion of running words, implying that speeches as other monologic forms, especially in formal debating contexts, can have significant lexical density.

Within political discourse, nominalization is a prevalent linguistic feature in monologic, which is the creation of a noun from a verb or an adjective, contributing to the density and formality of

the text; for instance, "*globalize*" can be nominalized as "*globalization*." Some scholars (Billig, 2008; Halliday, 1994; Kazemian & Hashemi, 2014) claim that the role of nominalization in contributing to the density and abstraction of informational texts, which is particularly pertinent in monologic political and academic discourses that often necessitate a formal and objective tone, and less common in dialogic discourse as it makes the conversation sound formal and less dynamic. Other scholars (Chilton, 2004; Fairclough, 2001; Fowler et al., 1979; Van Dijk, 1998) have investigated the functionality of nominalization in political discourse text, finding it more prevalent in monologic discourse.

Dialogic discourse often leverages modal verbs such as "*might*" and "*could*" to navigate various perspectives, express possibilities, and maintain politeness, a phenomenon confirmed by linguistics experts (Hyland, 2005). Concurrently, Austin (1962) and Searle (1969) delve into how modal verbs in dialogic contexts perform diverse actions, from making requests to suggesting possibilities. Furthermore, Tannen (1981) highlights the instrumental role of modal verbs in dialogic exchanges, softening directives, and exploring probabilities to maintain a balanced and cooperative interactional space. Together, these scholars underscore the pronounced presence of modal verbs in conversational discourse, shaping interactional dynamics and facilitating nuanced communication.

Quirk et al. (1985) noted the scarcity of interjections in formal, monologic discourse to sustain a formal and structured communication style. In contrast, Ameka (1992), Dingemanse (2023,2), and Tottie (1991) illustrated how conversational spoken sequences generously employ interjections, serving various *pragmatic* roles, managing interactions and concisely expressing attitudes and emotional reactions, enhancing dialogic interactions' dynamic and expressive nature.

Liu (2022) found that passive voice accounts for approximately 2% to 20% of English political speeches, with an average of 10%, predominantly used to state facts and emphasize opinions. Also, Heeman et al. (1998) studied discourse marker (words like "*so*", "*because*", or "*however*") usage in spontaneous speech using machine learning (ML), suggesting they help signal discourse structure and speaker intentions in dialogic contexts, including political communication. Finally, personal pronouns like "*you*", "*we*", and "*I*" are used in discourse to create personal references to the speaker, addressee, or others, being more useful in various communicative strategies, including those employed in dialogic exchanges (Kitagawa & Lehrer, 1990). Personal pronouns are crucial in thematic control, acknowledgment, and distance management of communication, particularly in dialogic communication, where they are instrumental in managing conversational dynamics and facilitating the coherence of discourse (Du Bois, 2007,1).

These studies collectively demonstrate the feasibility of operationalizing these linguistic features computationally in a text classification task for speeches and interviews, as shown in Table 1.

The need to classify texts into monologic and dialogic categories can be found in conversation analysis, the research area in linguistics that examines social interaction and its structure. In the pioneering work in this area, Sacks et al. (1974) developed a model for turn-taking in conversations. However, these methods relied heavily on analysis that involved intense manual labor—a standard practice at the time—making them infeasible for handling large volumes of data.

More recently, the field of NLP has witnessed a significant evolution with modern advancements encapsulated in libraries like spaCy or NLTK (Bird et al., 2009), which now come endowed with powerful pre-trained models. These contemporary tools present an opportunity to handle and

Table 1: Comparative of Linguistic Features of Speeches and Interviews Discourse According to Literature Review.

Feature	Speeches (monologic)	Interviews (dialogic)
Sentence Length	Long	Short
Word Length	Long	Short
Sentence Complexity	High	Low
Passive Voice Frequency	High	Low
Lexical Word Frequency	High	Low
Nominalization Frequency	High	Low
Personal Pronoun Frequency	Low	High
Interjection Frequency	Low	High
Modal Verb Frequency	Low	High
Discourse Marker Frequency	Low	High

measure linguistic features of text datasets and unravel the "*why*" behind decisions, adding a layer of transparency so crucial in many application domains.

Understanding how linguistic features at different levels—such as morphology, syntax, semantics, and pragmatics—impact decision-making in NLP models is crucial (Jurafsky & Martin, 2025). Moreover, the explainability of linguistic features facilitates model debugging, identifies and mitigates biases, and ensures that the system adheres to ethical and legal standards if necessary. Recent NLP explainability research emphasizes the synergy between linguistic features and neural models for enhancing transparency. Studies by Jumelet & Zuidema (2023), alongside Zhang et al. (2019), explore neural networks' feature interactions and interpretable modeling, highlighting the models' grasp of grammatical structures. Zafar et al. (2021a) assess neural text classifiers' interpretive reliability, finding unexpected behaviors, while Li (2022) and Yin & Neubig (2022) investigate contrastive explanations and the mutual benefits of theoretical linguistics and neural models, respectively. More specifically, studies have used SHAP to explain linguistic aspects of text classification (Vanni et al., 2023; Xiaomao et al., 2019; Zafar et al., 2021b; Zhao et al., 2020). Despite efforts in using SHAP for explaining NLP model predictions at the sentence level (Mosca et al., 2022), SHAP's strength lies in its ability to dissect model predictions to the token level, offering a finer granularity limiting sentence-level explainability.

Studies found that anonymization, the removal or alteration of personal data within the text to prevent individual identification, besides safeguarding privacy, often shifts the focus of neural models towards leveraging generic linguistic features rather than relying on specific semantic or contextual information, potentially improving model generalization (Lee et al., 2021; Nikolaidis et al., 2021).

Finally, studies demonstrated that BERT models rely on both semantic and syntactic levels when processing text. Tenney et al. (2019) explored how BERT captures linguistic information across its neural network, finding that its architecture internally follows the traditional and interpretable NLP pipeline, with parts responsible for specific linguistic tasks, from syntactic relationships on the lower level to semantic understanding as the layers go higher. Further research supports this

split specialization of BERT at those two levels of linguistic understanding (Clark et al., 2019; Coenen et al., 2019; Htut et al., 2019; Jawahar et al., 2019; Li et al., 2020; Liu et al., 2019; Michel et al., 2019; Rogers et al., 2020).

3. Models

In this section, we describe the datasets and their preparation, including data collection, cleaning, and annotation (Section 3.1). We then explain how we set up and fine-tuned both feature-based and BERT models for the classification task (Section 3.2). Datasets, models, and code are available in our GitHub repository. Along the section, we introduce:

- three datasets: (1) *Speech-vs-Interview-Feat-Dataset*, (to simplify, *FeatDataset*), containing the measured linguistic features extracted from a collection of 1,089 speeches and interviews, (2) *Speech-vs-Interview-Dataset*, (to simplify, *TextDataset*), a collection of 4,670 segments split using a sliding window mechanism and interviews, and (3) *Speech-vs-Interview-Dataset-Anonym*, (to simplify, *TextDatasetAnonym*) the NER-anonymized version of TextDataset; and

- four models: (1) an LRBM built on spaCy, a rule-based model to extract linguistic features from the text, (2) *LogRegFeat*, an LR model fit with FeatDataset, (3) *Speech-vs-Interview-Classification-BERT* (to simplify, *BERT1*), a BERT model fine-tuned with TextDataset, and (4) *Speech-vs-Interview-Classification-BERT-Anonym* (to simplify, *BERT2*), a BERT model fine-tuned with TextDatasetAnonym.

3.1. Datasets

We parsed massively audio/transcribed public political discourses using an ad hoc web-crawler tool (Reyes, 2023a) from targeted websites, predominantly from presidents and vice-presidents (78%), and a smaller fraction from other political figures and government officials (22%). In a subsequent manual selection process, we gathered speeches (involving only one person) and interviews (involving two or more participants). The selection criteria we followed, aligning with the gold standard requirements, to add discourse texts to the dataset included:

- *Domain*: The discourse texts belonged exclusively to the political scene within the United States, with the spoken English being American.

- *Representativity*: The discourse texts were perfect examples of speeches and interviews. Discourses that required minor editing to fit into a perfect example of the targeted classes were accepted.

- *Length*: The minimum number of tokens per discourse was 450, with no limit on the maximum number.

The data labeling task was partially self-reported but primarily determined through human annotation, ensuring a methodical approach to data categorization and analysis. The data was

labeled through three rounds of annotation, with seven annotators classifying samples as "*speech*" or "*interview*" independently using an ad hoc web application (Reyes, 2023b) and following strict guidelines (see Appendix B), including manual cleaning and anonymization. This process included the removal of timestamps in different formats like "(00:02):", more prevalent in interviews), speaker labels ("Donald Trump", "DONALD TRUMP:", "President Trump:", or "DT:"), surrounding text not integral to the discourse (including publication date, headline, subheadline, summary, etc.), and the anonymization of speakers and cross-references within the dialogues. In the case of hybrid discourses (town halls or conferences with a speech and questions and answers round later), the speech or interview part was removed, according to which discourse class was more dominant. A cleaning procedure was implemented to ensure data quality, including expanding contractions, removing URLs, Unicode symbols, speaker labels, applauses, cheers, bracket annotations, timestamps, and any contextual data. The spaCy library with the pre-trained *Transformer model* (*en_core_web_trf*) model and the "*BertTokenizer*" tokenizer from *Hugging Face*'s *Transformers* library were used for text preprocessing. With this process, we ensured that the discourse texts from both speech and interview classes were stripped of any identifiable information or patterns that could potentially bias the model's classification decisions.

The *Inter-Annotator Agreement* (*IAA*) analysis achieved a *Fleiss's Kappa* score of 0.956 on the classification facet (speech/interview), implemented using *statsmodels* (Seabold & Perktold, 2010). In the final round, we introduced an anonymous review, where two annotators were unaware of initial classifications, further reducing bias. An additional curator established the gold standard in case of disagreements. 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.

From the human perspective, the annotation task was straightforward because (1) it is moderately easy to distinguish if a discourse is a speech or an interview mostly by simple visual inspection or by the title, and (2) the annotation tool was built to facilitate the specific task, including cleaning and anonymization, demanding less cognitive load than other annotation tasks. However, from the machine's perspective, once the class-specific information was removed from the text, leaving only discourse-level content, the classification task became more complex and required fine-grained linguistic analysis. To ensure the model could analyze discourse as a whole, we preserved the dialogue exchange of all parties in the interview class as a cohesive unit and did not selectively remove dialog segments, such as the interviewers.

This initial corpus comprised 1,089 political discourse texts representing a wide range of speeches and interviews covering most of the political issues in the U.S.A., collected from multiple websites attempting to create a varied sampling of political discourse texts by different American politicians: *The American Presidency Project* (1,031), *Rev.com* (36), *National Archives and Records Administration* (6), *United States Senate* (5), *ABC News* (3), *NPR* (1), *United States House of Representatives* (5), *Cleveland.com* (1), and *The White House* (1). To create FeatDataset, we extracted numerical representations of its linguistic features using the LRBM. We used our understanding of the ten linguistic features previously studied (Table 1) as a foundational

Figure 1: Examples in the Speech-vs-Interview-Dataset (TextDataset) and Speech-vs-Interview-Dataset-Anonym (TextDatasetAnonym) datasets.

theoretical framework to build the LRBM. (We return to this issue in the experimental setup description.)

The datasets TextDataset and TextDatasetAnonym were built from the initial corpus using the *sliding window approach* by segmenting speeches and interviews in text sequences that do not exceed the model's fixed maximum input capacity of 512 tokens. TextDatasetAnonym was processed using spaCy's Transformer model, where all named entities were anonymized automatically and replaced with placeholders as in Figure 1: "PERSON", "ORG", "NORP", "TIME", "DATE", "CARDINAL", "MONEY", "FAC", "QUANTITY", "PERCENT", and "GPE". We selected Named Entity Recognition (NER) for anonymization because it is a well-established method, and its implementation can be efficiently automated using spaCy. The main goal of this procedure was to minimize the contextual information, such that the BERT models focus on linguistic structures and not on the context that named entities give.

After applying the sliding window technique and NER anonymization to the initial corpus of 1,089 texts, the dataset was balanced by downsampling the majority class via a random sampling process. This resulted in a final dataset of 4,670 datapoints (2,335 per class), as seen in Table 2. The random sampling process excluded examples with fewer than 450 tokens. The final source distribution for TextDataset and TextDatasetAnonym was as follows: The American Presidency Project (4,091), Rev.com (405), National Archives and Records Administration (70), United States Senate (27), ABC News (25), NPR (22), United States House of Representatives (12), Cleveland.com (10), and The White House (8).

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. The speakers' gender distribution was highly skewed, featuring 98.5% males and only 1.5% females.

Table 2: Datasheets of Datasets for Speech and Interview Classification.

	FeatDataset	**TextDataset**	**TextDatasetAnonym**
Name	*Speech-vs-Interview-Feat-Dataset*	*Speech-vs-Interview-Dataset*	*Speech-vs-Interview-Dataset-Anonym*
Instances	Speeches and Interviews by American politicians	Segments of political discourse texts by American politicians	Segments of political discourse texts by American politicians
Classes (*)	• Speech (s) • Interview (i)	• Speech (s) • Interview (i)	• Speech (s) • Interview (i)
Number of Instances	1,089 (537 s/552 i)	4,670 (2,335 s/2,335 i)	4,670 (2,335 s/2,335 i)
Instance Length	Between 468 to 24,604 tokens	Between 450 to 512 tokens	Between 450 to 512 tokens
Labels	• "speech" • "interview"	• "speech" • "interview"	• "speech" • "interview"
Splits/Instances	• Train: 870 (80%) • Test: 219 (20%)	• Train: 3,736 (80%) • Validation: 466 (10%) • Test: 468 (10%)	• Train: 3,736 (80%) • Validation: 466 (10%) • Test: 468 (10%)
Stratification	• Train: 429 s and 441 i • Test: 108 s and 111 i	• Train: 1868 s and 1868 i • Validation: 233 s and 233 i • Test: 234 s and 234 i	• Train: 1868 s and 1868 i • Validation: 233 s and 233 i • Test: 234 s and 234 i
Metadata	• title (document) • source_url • politician_name • gender • publication_date	• title (document) • source_url • politician_name • gender • publication_date	• title (document) • source_url • politician_name • gender • publication_date
Data Period	1939–2023	1939–2023	1939–2023

3.2. Experimental Set-up

Based on the existing literature on the linguistic features of speeches and interviews, we developed the LRBM using a set of rules to extract the measurement of the frequency of ten specific features, relying on the linguistic capabilities provided by spaCy, including *sentence segmentation*, *part-of-speech tagging*, *statistical and rule-based morphology*, *lemmatization*, *pattern matchers*, and *dependency parsing*. We chose spaCy for its strong linguistic capabilities, which allow for the interpretable application of linguistic theories in the collected transcribed text through a unified and centralized framework. This feature engineering modeled the LRBM implementing count-based methods to measure feature occurrence and frequency and opted for the options that were primarily more human-understandable and, secondarily, easier to implement.

- *Sentence Length*: The count of words (tokens) in a sentence, excluding punctuation.
- *Word Length*: The count of the number of characters in words (tokens), excluding those punctuation.

- *Sentence Complexity*: The count of the number of adverbial clauses per sentence.
- *Passive Voice Frequency*: The count of the occurrences of passive voice constructions within sentences.
- *Lexical Word Frequency*: The count of lexical words (nouns, verbs, adjectives, adverbs) within sentences.
- *Nominalization Frequency*: The count of nominalizations per sentence was measured, highlighting the use of noun forms derived from verbs or adjectives using a lexicon of suffixes often used in nominalization (see Appendix A).
- *Personal Pronoun Frequency*: The count of the instances of personal pronouns in each sentence, according to a lexicon of personal pronouns excluding reflexive pronouns (see Appendix A).
- *Interjection Frequency*: The count of how often interjections appear within sentences, according to a lexicon of interjections (see Appendix A). (We return to this issue in more detail below.)
- *Modal Verb Frequency*: The count of modal verb occurrences in sentences.
- *Discourse Marker Frequency*: The count of discourse markers that appeared in sentences according to a lexicon of adverbs and conjunctions commonly used in discourse texts (see Appendix A) to convey relationships and connections, which usually signal transitions, emphasize information, indicate contrast, introduce examples, express cause and effect, and similar.

Interjections are determined in spaCy primarily by a combination of a statistical model and lexical attributes. spaCy's models consider the surrounding words and syntax, such that if a token appears in a position where an interjection would typically be found, the model is likely to tag it. But also, some words are marked as interjections using spaCy's lexical data (e.g., "*wow*", "*oops*"); if a token is recognized from this set, it may automatically be tagged as it. The lexicon of interjection was collected by applying spaCy's "INTJ" POS tag to each token in the test datasets. During this automatic process, some tokens that are not strictly interjections were included in the collection, but we decided to retain them after observing that they can function as interjections in natural language (e.g., "*see*", "*thank*", "*kid*").

Although *question frequency* appeared promising, given its relevance to interviews, we ultimately excluded it as a feature due to methodological limitations. spaCy does not provide a native, reliable method for detecting question structures, and a simple count of question marks does not reflect the turn-taking dynamic in discourse, especially given the prevalence of rhetorical questions in political discourse and inconsistent punctuation in transcripts. Our preliminary attempts to measure this feature highlighted the ambiguity and syntactic complexity of questions in political discourse, which would require more advanced parsing to capture accurately.

We undertook a comprehensive preprocessing and exploratory data analysis (EDA) phase to understand the feature dataset's characteristics in both classes and to inform our data preparation

decisions. Initial observations of the distribution of each feature in both classes were conducted through histogram, boxplot, and scatter plot visualizations. We employed an LR model (LogRegFeat) to establish a foundational baseline in our study due to its interpretability and efficacy in handling binary classification tasks, serving not only as a benchmark for performance comparison but also as a tool for understanding the impact and importance of each involved feature, paving the way for a deeper investigation into BERT's explainability and its ability to generalize.

We employed TextDataset and TextDatasetAnonym, divided into training (80%), validation (10%), and testing (10%) subsets, to fine-tune the BERT models on a *CUDA-enabled GPU* (*NVIDIA GeForce GTX 1080*), utilizing the "*bert-base-uncased*" pre-trained model variant 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—a process that took weeks to find the highest accuracy in both models. We monitored the *training and validation losses* closely, employing the early-stop strategy when the training loss ceased to decrease, thereby preventing overfitting. For *BERT1*, we used the following metrics: *learning rate*, 1.2465928099530177e-05; *batch size*, 16; *warm-up steps*, 369; *number of epochs*, 4; and *seed*, 42. For *BERT2*, we used the following metrics: *learning rate*, 2.1710126259258467e-05; *batch size*, 16; *warm-up steps*, 896; *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).

To reduce the risk of our model relying on thematic or context-specific cues rather than structural/linguistic features, we used a chi-square test to identify the terms most strongly correlated with each class (i.e., class-indicative terms). We then replaced the top 50 of these highly class-indicative terms with neutral placeholders, thereby minimizing the influence of thematic content on the classification process. After pruning unimportant features, we examined the influence of linguistic features on BERT's decision-making through SHAP analysis for class-wide explanation. Our approach focused on a token-level analysis. Our XAI setup followed the same procedure for both datasets, TextDataset and TextDatasetAnonym: (1) we split the test dataset into the two classes, with 234 examples per class; (2) we submitted each example to the BERT model through a SHAP explainer, generating values for each token (word) according to their importance in the model's predictions; and (3) we aggregated mean absolute SHAP values across classes to each feature using the LRBM to map features according to lexicons. Figure 2 shows a high-level diagram of the explanation architecture.

Feature data with positive skew were normalized with a log transformation, and a *Bonferroni* correction was applied to avoid false-positive results (*Type I* errors) when performing multiple statistical tests simultaneously. Subsequent independent two-sample (*Welch*) t-tests compared the mean absolute SHAP values, identifying linguistic features with significantly different SHAP values between discourse types, consequently revealing BERT's reliance on linguistic structure versus thematic content, indicating its generalization capability and elucidating the explainability of its classification decisions.

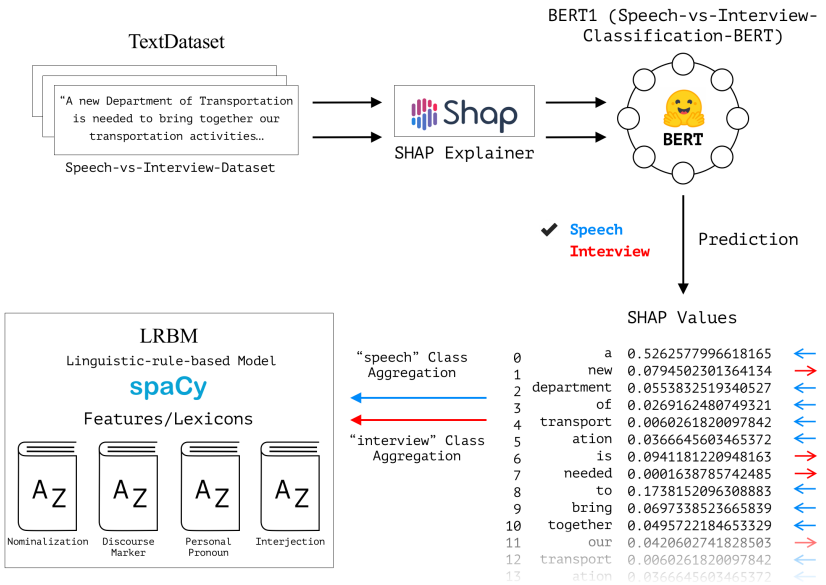

Figure 2: XAI Setup using SHAP on BERT1, TextDataset, and the LRBM for Speech-vs-Interview Classification.

4. Results

This section presents the results of our analysis, highlighting key linguistic differences between political speeches and interviews (Section 4.1). We then discuss feature importance in classification (Section 4.2) and evaluate the performance of different models (Section 4.3). Lastly, we explore bias mitigation (Section 4.4) and interpretability through SHAP analysis (Section 4.5).

4.1. Linguistic Differences Between Speeches and Interviews

As shown in Table 3, the results from the descriptive statistics and EDA stage highlight significant linguistic differences between political speeches and interviews.

The mean sentence length was approximately 17 words for both speech and interview formats. Speeches exhibited a more comprehensive range of sentence lengths compared to interviews (278), pointing to a mix of very short and very long statements in some speeches. The average word length in speeches was 4.39, slightly longer than the 4.25 observed in interviews, suggesting that speeches may employ more polysyllabic words. Sentence complexity scores differed, with speeches showing a mean of 1.36 and interviews a mean of 1.63, indicating that interview sentences can be more syntactically complex.

Passive voice frequency was low in both formats, with interviews displaying a slightly higher mean, implying that both remain largely active-voice, but interviews occasionally incorporate

Table 3: Statistics Summary for the Speech and Interview Classes.

Feature	Mean	Range	SD	Variance	Skewness	Kurtosis
Speech Class						
Sentence Length	17.20	278	13.06	170.51	2.13	10.58
Word Length	4.39	18	2.41	5.82	1.12	1.07
Sentence Complexity	1.36	27	1.63	2.65	2.06	7.77
Passive Voice Freq	0.11	8	0.36	0.13	**3.83**	**21.27**
Lexical Word Freq	7.40	107	5.93	35.12	2.05	9.20
Nominalization Freq	0.80	19	1.19	1.42	2.36	9.54
Personal Pronoun Freq	1.76	19	1.59	2.53	1.79	5.96
Interjection Freq	0.03	4	0.20	0.04	**7.23**	**66.27**
Modal Verb Freq	0.30	8	0.61	0.38	2.58	9.44
Discourse Marker Freq	0.85	12	1.17	1.37	2.05	6.56
Interview Class						
Sentence Length	17.71	187	14.26	203.46	1.84	5.70
Word Length	4.25	19	2.38	5.65	1.19	1.35
Sentence Complexity	1.63	26	1.91	3.67	1.92	5.85
Passive Voice Freq	0.12	7	0.37	0.14	**3.58**	**16.21**
Lexical Word Freq	7.24	83	6.19	38.29	1.91	6.23
Nominalization Freq	0.77	14	1.11	1.22	2.10	6.69
Personal Pronoun Freq	1.90	21	1.72	2.95	1.75	5.42
Interjection Freq	0.11	6	0.35	0.12	**3.60**	**17.63**
Modal Verb Freq	0.33	10	0.65	0.42	2.49	8.75
Discourse Marker Freq	1.00	16	1.28	1.63	1.88	5.33

more passive constructions. Skewness and kurtosis values for passive voice frequency were 3.83 and 21.27 for speeches, respectively, and 3.58 and 16.21 for interviews, respectively, indicating that although the overall average use of passive voice is low, a few speeches show relatively high usage. Lexical word frequency measures indicated a broad range in both speech and interview formats, with maximum values slightly higher in speeches, pointing to certain speeches featuring a more diverse or dense vocabulary. The mean values for nominalization were 0.80 in speeches and 0.74 in interviews, reflecting slightly more abstract word forms in speeches overall.

Personal pronoun frequency was more frequent in interviews, with a mean of 1.90, compared to 1.76 in speeches, possibly highlighting the interactive nature of interviews where participants refer to themselves or each other more often. Interjection frequency was higher in interviews, with a range of up to 6 and a mean of 0.11. Speeches had skewness and kurtosis values for interjection frequency at 7.23 and 66.27, respectively, meaning that while most speeches rarely use interjections, a small number include them at very high rates, producing a long-tail distribution. Modal verb frequency was relatively low in both formats, with mean values of 0.30 for speeches and 0.33 for interviews, indicating only minimal differences in the use of possibility or necessity

expressions. Discourse marker frequency was more used in interviews, with a mean of 1.00, compared to 0.85 in speeches. High skewness and kurtosis values were noted for many features in the speech data, suggesting that although most speeches converge around typical usage rates, some deviate substantially, forming a long-tail distribution.

Figure 3 presents histograms illustrating the distributions of interjection frequencies within each class. The interview histogram displays a distribution that differs notably from speeches. Specifically, the interview distribution shows a broader spread of interjection frequencies, whereas a sharp peak and a pronounced long tail characterize the speech distribution. This higher kurtosis implies that typical speeches tend to cluster around low interjection frequency, but a few have very high frequency, causing extreme peaks in the distribution.

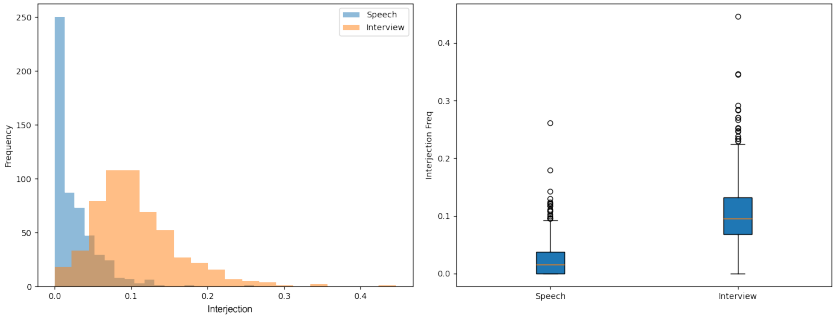

Figure 3: Histogram and Box Plot for Interjection Frequency in Speeches and Interviews.

To confirm the descriptive statistics (Table 3), Welch independent-samples t-tests were conducted. Results showed that interviews featured a significantly higher frequency of interjections ($M = 0.11$, $SD = 0.35$) than speeches ($M = 0.03$, $SD = 0.20$), $t(909.5) = -4.51$, $p < .001$. Similarly, interviews showed significantly higher sentence complexity ($M = 1.63$, $SD = 1.91$) compared to speeches ($M = 1.36$, $SD = 1.63$), $t(1083.2) = -2.71$, $p = .007$, and a higher frequency of discourse markers ($M = 1.00$, $SD = 1.28$) than speeches ($M = 0.85$, $SD = 1.17$), $t(1084.9) = -2.26$, $p = .024$. Other feature differences between the two classes were not statistically significant.

Correlation analyses were conducted to explore relationships among various linguistic features within the speech and interview classes. Figure 4 illustrates the correlation findings for the speech class. A strong correlation was observed between sentence length and lexical word frequency ($r = 0.95$), indicating a relationship where longer speeches are associated with a broader vocabulary. The correlation coefficient between sentence complexity and personal pronoun frequency was significant at $r = 0.67$. A notable correlation was also observed between the use of discourse markers and both sentence length ($r = 0.70$) and lexical diversity ($r = 0.62$). Additionally, a negative correlation between personal pronoun frequency and word length ($r = -0.43$) was recorded.

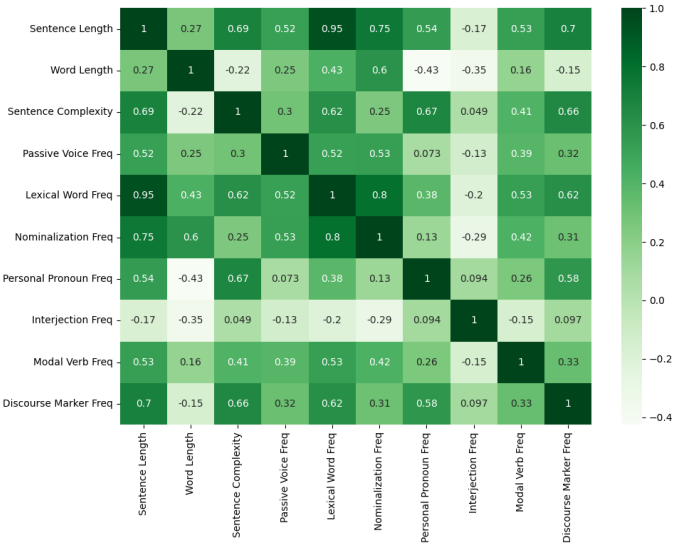

Figure 4: Correlation Analysis for the Speech Class.

In the interview class, as depicted in Figure 5, the analysis also revealed a strong correlation between sentence length and lexical word frequency ($r = 0.98$). Further correlations were identified between sentence complexity and personal pronoun frequency ($r = 0.75$) and between discourse markers and personal pronoun frequency ($r = 0.75$), indicating characteristics of the interactive nature of interviews.

For both speeches and interviews, correlations were explored between sentence length and sentence complexity, with coefficients of $r = 0.69$ for speeches and $r = 0.88$ for interviews, suggesting a trend where longer sentences are more complex in both formats. Negative correlations were found between interjection frequency and sentence length ($r = -0.17$ for both speeches and interviews) and between interjection frequency and word length ($r = -0.35$ for speeches and $r = -0.30$ for interviews), highlighting a trend towards less frequent use of interjections in more formally structured discourse.

We identified redundant features via correlation analysis and observation of dependencies among features rooted in established linguistic principles. For instance, a sentence with more lexical words (nouns, verbs, adjectives, and adverbs) tends to be longer because lexical words carry the core meanings and concepts, as opposed to function words, which primarily serve grammatical purposes and are usually shorter. Likewise, a sentence with more nominalizations will contain a higher count of lexical words, as nominalizations—nouns derived from verbs or adjectives—contribute to the overall lexical density of the sentence. Therefore, we pruned sentence length, sentence complexity, and lexical word frequency. Given that SHAP explanations focus on

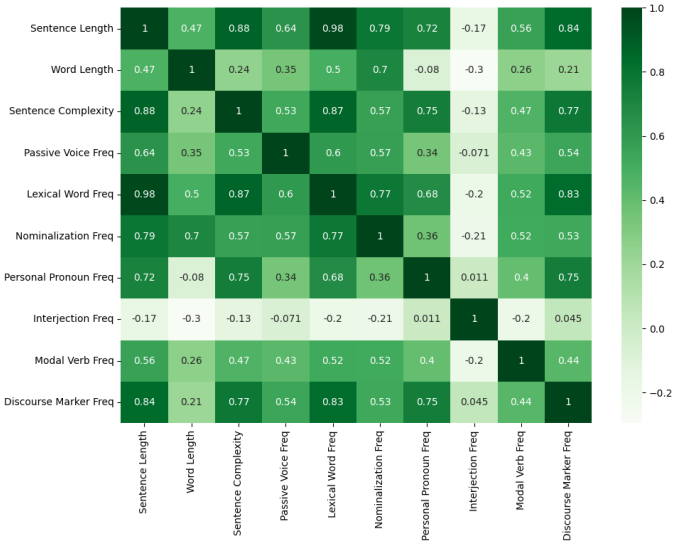

Figure 5: Correlation Analysis for the Interview Class.

the word or *subword* level, we also pruned passive voice due to their impracticality in evaluating them at the token level. Finally, we pruned word length from LogRegFeat because it acts more as an implicit feature rather than a direct linguistic indicator. Since BERT already captures word length implicitly through tokenization—where longer words are often split into multiple subword tokens—explicitly including it in LogRegFeat would provide limited additional value. (We return to this issue in the Discussion section.)

4.2. Feature Importance

Several significant findings were observed in the logistic regression analysis of LogRegFeat, as reported in Table 4. Interjection frequency emerged as a predictor, displaying a significant negative relationship with the likelihood of a discourse being classified as a speech. The analysis revealed a coefficient (β) of -22.330, a standard error (SE_β) of 1.575, and a z-value of -14.176, all culminating in a p-value of less than .001, resulting in an odds ratio (e^β) of approximately 2.00×10^{-10}, suggesting that a higher frequency of interjections is more characteristic of interviews than speeches. Similarly, modal verb frequency was found to significantly negatively affect speech classification, as indicated by a β of -4.126, an SE_β of 0.932, a z-value of -4.426, and a p-value of less than .001, with e^β of 1.61×10^{-2} further highlights the tendency for a lower frequency of modal verbs in speeches compared to interviews.

Table 4: LogRegFeat Results for the Classification of Speeches and Interviews.

Feature	β	SE_β	z	p	e^β	Effect
const (intercept)	2.936	0.457	6.426	$< .001$	18.8	
Interjection	-22.330	1.575	-14.176	$< .001$	$2.00 \cdot 10^{-10}$	Strong negative
Modal Verb	-4.126	0.932	-4.426	$< .001$	0.0161	Significant negative
Discourse Marker	0.518	1.247	0.415	**.678**	1.68	Significant positive
Personal Pronoun	1.079	1.095	0.986	**.324**	2.94	Not significant
Nominalization	2.569	1.159	2.217	.027	13.0	Not significant

Discourse marker frequency, however, did not show a significant effect on the classification outcome, with a β of 0.518, an SE_β of 1.247, a z-value of .415, and a p-value of 0.678. The e^β stood at 1.68, suggesting a minimal impact on differentiating between speeches and interviews. Despite a positive β of 1.079 and an e^β of 2.94, personal pronoun frequency did not reach statistical significance, with an SE_β of 1.095, a z-value of 0.986, and a p-value of 0.324. This outcome suggests that while there might be a trend towards higher personal pronoun frequency in speeches, the evidence is not strong enough to confirm a significant effect. Finally, the frequency of nominalizations showed a positive association with speech classification, as evidenced by a β of 2.569, an SE_β of 1.159, a z-value of 2.217, and a p-value of .027, resulting in an e^β of 13.0, indicating that discourses with a higher occurrence of nominalizations are more likely to be classified as speeches.

4.3. Models Performance

As seen in Table 5, LogRegFeat, serving as a baseline, exhibits a significant accuracy, precision, recall, and F1 score of 0.890, with an AUC-ROC score of 0.953, indicating a high level of performance in binary classification speeches and interviews based on linguistic features. While it shows a slight preference in recall for interviews over speeches, it remains effective and reliable for this classification task. On the other hand, BERT1 shows a remarkable improvement in all metrics, achieving accuracy, precision, recall, and F1 score of 0.981, alongside an AUC-ROC score of 0.993. BERT2 also performs impressively, with accuracy, precision, recall, and F1 score of 0.974, and an AUC-ROC score of 0.995. The slight decrease in accuracy, precision, recall, and F1 score of BERT2, compared to BERT1, indicates that while the model can still effectively classify speeches and interviews without specific names and identifiers, it relies to some extent on these elements for achieving the highest performance. However, the increase in the AUC-ROC score suggests that BERT2 is slightly more effective in distinguishing between classes at various threshold settings, possibly due to its focus on linguistic structure over thematic context.

The performance of both BERT models reinforces their capabilities to understand and analyze political communication, offering insights into their capabilities and limitations in handling complex linguistic patterns. Specifically, the slight performance dip in the BERT2 and its higher

[1] s = speech, i = interview. Across-class metrics are macro and class-wise metrics are not averaged

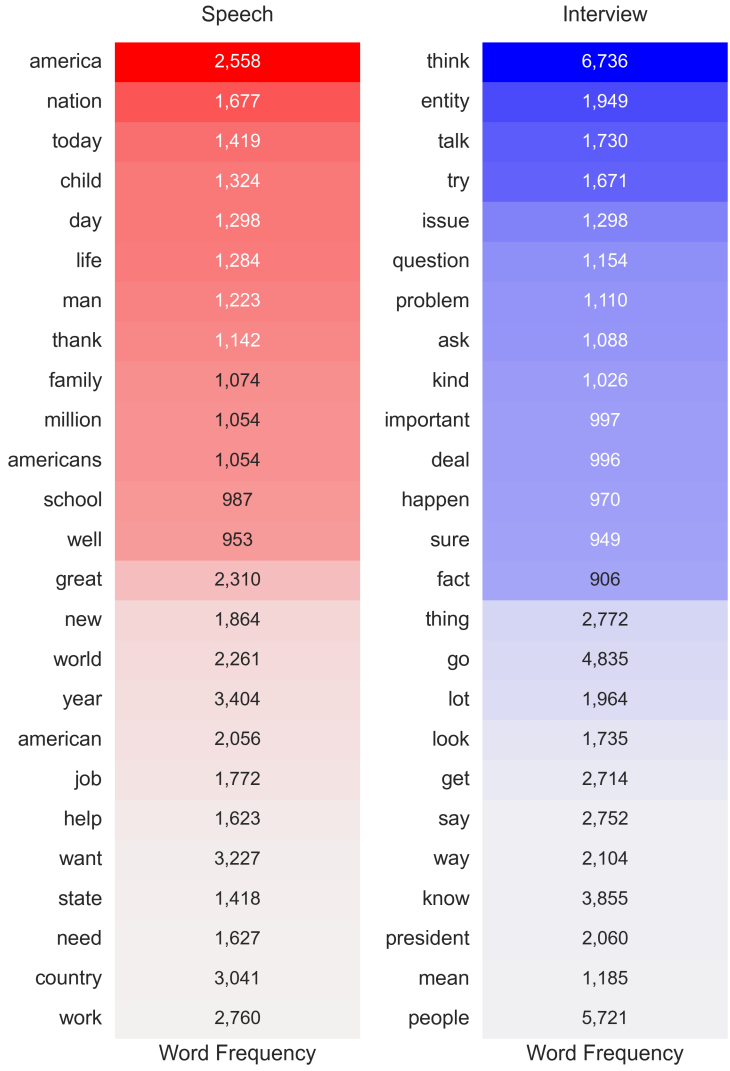

Figure 6: Heatmap of Word Frequency Bias in Political Speeches and Interviews from Chi-Square Analysis. The first 50 (lowered and lemmatized) terms per class were ordered by chi-square coef-ficients, which show the number of term occurrences

Table 5: Summary of performance metrics of the LogRegFeat, BERT1 and BERT2 for classifying Speeches and Interviews.

Metric	LogRegFeat			BERT1			BERT2		
Accuracy		0.890			0.981			0.974	
Precision (macro)		0.890			0.981			0.974	
Recall (macro)		0.890			0.981			0.974	
F1 Score (macro)		0.890			0.981			0.974	
AUC-ROC		0.953			0.993			0.995	
		s	*i*		*s*	*i*		*s*	*i*
Confusion Matrix[1]	*s*	*96*	*16*	*s*	*231*	*3*	*s*	*227*	*7*
	i	*8*	*100*	*i*	*6*	*228*	*i*	*5*	*229*
Speech Class									
Precision		0.922			0.975			0.978	
Recall		0.856			0.987			0.970	
F1-score		0.888			0.981			0.974	
Interview Class									
Precision		0.862			0.987			0.970	
Recall		0.926			0.974			0.979	
F1-score		0.893			0.981			0.974	

AUC-ROC score (0.995) may suggest that anonymization compels the model to rely more on linguistic structures on thematic content—a hypothesis which we examine through SHAP analysis and worth validating in future research through error analysis or qualitative evaluation. Datasets and models are publicly available (Reyes, 2026).

4.4. Bias Mitigation

The chi-square analysis to identify terms with statistically significant biases towards either class in the three splits of TextDataset revealed pronounced biases for a wide range of terms, indicating notable differences in term usage that reflect the unique communicative dynamics of speeches and interviews. Although these terms can be considered important features in this classification task and not problematic bias, we mitigated their influence just to improve the experimental environment according to our explainability goals. As shown in Figure 6, significant biases were observed for terms such as "*think*" ($\chi^2 = 2772.379$, $p < 0.001$), which was predominantly used in the interview class, and "*america*" ($\chi^2 = 2649.001$, $p < 0.001$), which showed a preference for the speech class. Similarly, the term "*talk*" ($\chi^2 = 1707.404$, $p < 0.001$) was found to be more frequent in interviews, whereas "*nation*" ($\chi^2 = 1727.455$, $p < 0.001$) was more commonly associated with speeches. The analysis extended to a variety of other terms, with "*today*" ($\chi^2 = 1459.287$, $p < 0.001$) and "child" ($\chi^2 = 1360.737$, $p \approx 0$) signaling biases towards speeches, while terms like "*try*" ($\chi^2 = 1648.566$, $p < 0.001$) were more frequently used in interviews. The term "*entity*" is a special case because it is the placeholder of the names of

Table 6: Summary of t-tests for Speech and Interview Classes Based on SHAP values.

Model	Feature	t	p-value	Speech		Interview	
				M	*SD*	*M*	*SD*
BERT1	Nominalization Frequency	-4.2	$< .001$	-1.79	.20	-1.38	.30
	Discourse Marker Frequency	-4.9	$< .001$	-3.04	.33	-2.23	.53
	Personal Pronoun Frequency	-3.9	.001	-3.64	.21	-3.29	.26
	Interjection Frequency	-2.9	.008	-4.26	.78	-3.34	.91
BERT2	Nominalization Frequency	-0.7	0.516	-1.20	.32	-1.12	.29
	Discourse Marker Frequency	0.0	0.971	-2.38	.70	-2.39	.59
	Personal Pronoun Frequency	-2.4	0.026	-3.39	.13	-3.15	.32
	Interjection Frequency	-1.1	0.277	-3.60	1.01	-3.18	.87

the interviewer and interviewee "ENTITY", described above, to anonymize interviews where participants use their real names or titles. The described patterns illustrate the differences in term usage in speeches and interviews, with speeches perhaps focusing more on evoking nationalistic and familial sentiments, as indicated by the frequent use of "*nation*", "*child*", and "*family*", whereas interviews tend to prioritize discussion, reflection, and questioning, as seen with "*think*", "*question*", and "*talk*".

4.5. SHAP Analysis

From the t-tests, we observed a limitation in analyzing modal verb frequency since, in English grammar, modal verbs are only nine, creating a limitation in the aggregation of SHAP values. Therefore, we removed it from further analysis.

T-test results on SHAP values from BERT1 revealed that the four features serve as critical discriminators in the model's classification process (Table 6). The significant differences in how these features influence model predictions across the two discourse types suggest that each plays a vital role in enabling the model to recognize and differentiate between speeches and interviews. Plots in Figure 7 visually substantiate the statistical analysis, demonstrating that BERT1 does not treat all linguistic features equally, thus contributing to the explainability of the model in terms of feature importance.

However, the results from BERT2 show a different pattern, particularly for nominalization frequency and interjection frequency, where the p-values indicate a lack of statistical significance. This suggests that with anonymization, BERT2's ability to rely on these specific linguistic features diminishes, pointing towards a decrease in the model's sensitivity to certain linguistic structures when contextual clues are minimized. However, personal pronoun frequency showed a significant difference ($p = 0.026$), indicating its pivotal role in BERT2's ability to distinguish between the two types of discourse interpreted by SHAP. This contrasting result indicates that BERT2 has a reduced sensitivity to specific linguistic structures, and it may leverage other aspects of the data or rely on a more generalized understanding of the text to make its classifications.

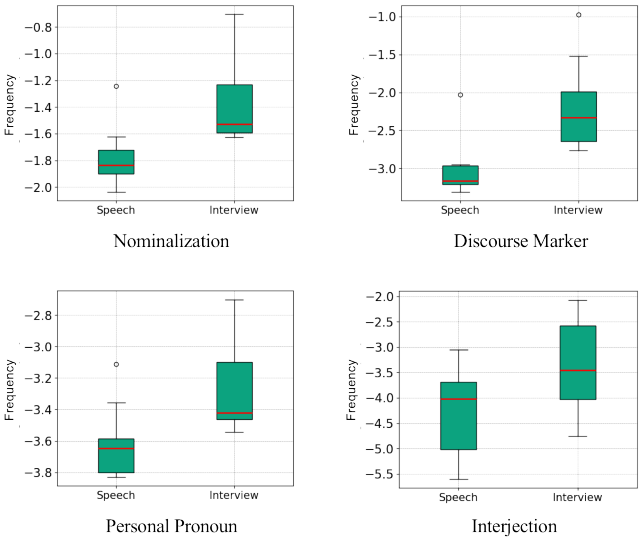

Figure 7: Comparison of features between Speech and Interview classes from SHAP values analyzed in BERT1.

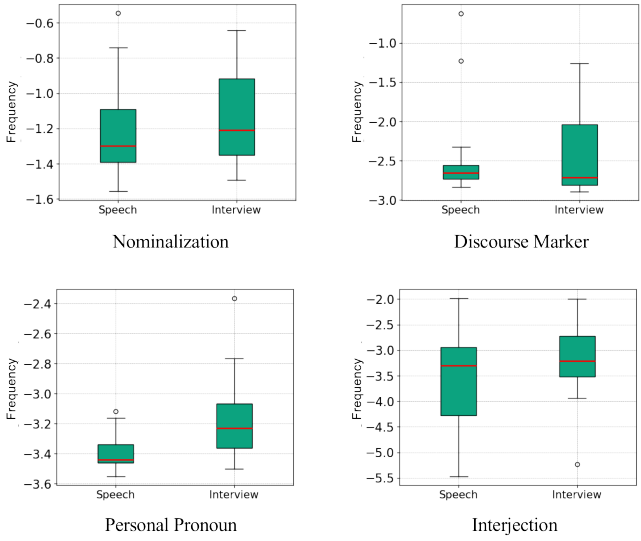

Figure 8: Comparison of features between Speech and Interview classes from SHAP values analyzed in BERT2.

Table 7: Alignment of Literature Review and Descriptive Statistic Analysis. The labels in the Statistics columns are based on a comparison of the mean values reported in Table 3.

Feature	Speech (monologic)		Interview (dialogic)		Alignment
	Literature	Statistics	Literature	Statistics	
Sentence Length	**Long**	**Short**	**Short**	**Long**	**No**
Word Length	Long	Long	Short	Short	Yes
Sentence Complexity	**High**	**Low**	**Low**	**High**	**No**
Passive Voice Frequency	**High**	**Low**	**Low**	**High**	**No**
Lexical Word Frequency	High	High	Low	Low	Yes
Nominalization Frequency	High	High	Low	Low	Yes
Personal Pronoun Frequency	Low	Low	High	High	Yes
Interjection Frequency	Low	Low	High	High	Yes
Modal Verb Frequency	Low	Low	High	High	Yes
Discourse Marker Frequency	Low	Low	High	High	Yes

Plots in Figure 8 suggest that anonymizing the dataset could lead to a more uniform distribution of SHAP values across features, implying that BERT2's reliance on these linguistic structures might be less pronounced when semantic and contextual identifiers are removed.

5. Discussion

The following analysis should be understood with a key methodological choice in mind: the interviews retain contributions from all participants (both interviewer(s) and interviewee(s)), which influences the metrics of the interview class. This chapter elaborates on the findings of our study, examining how linguistic features shaped the predictions of the introduced models (Section 5.1) and comparing BERT's behavior across anonymized and non-anonymized data, offering insights into our goals of AI transparency (Section 5.2).

5.1. Linguistic Features

As observed in Table 7, certain features in the analyzed political discourse texts did not align with the traditional linguistics of monologic and dialogic discourse. Conventionally, sentence length in monologic communication is expected to be longer due to a tendency towards elaboration and detailed explanation. Contrary to these expectations, our statistical analysis of Table 3 revealed a shorter average sentence length in speeches ($M = 17.20$ words) compared to interviews ($M = 17.71$ words). Notably, however, the range of sentence lengths was substantially wider in speeches ($R = 278$ words) than in interviews ($R = 187$ words). This discrepancy suggests that while speeches, on average, may utilize shorter sentences, they also exhibit a higher variability in sentence length, allowing succinct statements and extensive elaboration within the same discourse context. These findings may reflect a strategic use of sentence variety to maintain audience engagement or may be influenced by the specific corpus and contexts of the collected discourse texts. We also

observed a considerable correlation between sentence length and sentence complexity ($r = 0.69$), which is expected since the complexity of a sentence may be reflected in its length.

Literature on passive voice views it as a feature of formality and impersonality, expecting a higher prevalence in speeches. However, in political interviews, the frequency of passive voice ($M = 0.12$) is not markedly higher than in interviews ($M = 0.11$), and the kurtosis is substantially higher in speeches ($\kappa = 21.27$) compared to interviews ($\kappa = 16.21$), which signals that passive voice is used either very frequently or very infrequently. This discrepancy may be due to a rhetorical shift toward intentionally adopting the active voice to emphasize their personal involvement and efficacy. Compared to the interview class, speeches demonstrated greater feature variability, suggesting speeches' prepared and carefully orchestrated nature, where speakers consciously vary language to achieve desired rhetorical effects. These observations confirm that political discourse has its own linguistic norms that can reflect and deviate from general discourse patterns, depending on the rhetorical goals and context.

The correlation analysis also observed several significant relationships between linguistic features within speech and interview classes, like the strong positive correlations between the use of personal pronouns that correlated positively with sentence complexity in both communication modes ($r = 0.67$ for speech; $r = 0.75$ for interview), suggesting a link between the personalization of language and the construction of more complex sentences, being more noticeable in interviews, where interpersonal interaction, expression of ideas, the need of building relations is necessary for effective communication.

EDA highlighted the consistent use of interjections in interviews ($M = 0.11$), something aligned with the literature, pointing out that the conversational nature of interviews signals engagement, agreement, or other reactions within an interactive setting. The histogram for speeches in Figure 3 shows a highly skewed distribution with a sharp peak and long tail, indicating that interjections are generally infrequent in speeches but may occur in bursts or be highly pronounced when they do occur. This observation suggests that when interjections are used in speeches, they are likely very intentional.

As seen in Table 4, LogRegFeat found five features that influence the classification of discourse as speech or interview. The importance of interjection frequency ($\beta = -22.330$) and modal verbs ($\beta = -4.126$) in predicting the interview class underscores the greater reliance on informal language elements, spontaneous reactions, politeness, and softening requests, uncertainty in answers, need for conveying possibility, and other discursive strategies. Oppositely, the more formal and lexically diverse language in speeches, where the reported strength of nominalization ($\beta = 2.569$) and the subtle presence of discourse marker frequency ($\beta = 0.518$) indicated an association with a wide lexical variety, suggesting that speeches convey messages employing a broader vocabulary to enhance their impact and clarity. The higher use of personal pronoun frequency ($\beta = 1.079$) in speeches suggests the speaker's attempt to connect personally with the audience or express personal opinions or experiences. Nevertheless, the p-values of discourse marker frequency ($p = 0.678$) and personal pronoun frequency ($p = 0.324$) indicate that their contribution is not statistically significant or that the rules to detect and parse these features have limitations.

SpaCy's efficiency and effectiveness in capturing the linguistic features pertinent to our investigation are validated by aligning operationalized features with our literature review and expec-

tations. The explainable rules built with spaCy for parsing, identifying, and quantifying these features provided a reliable foundation for our analysis; hence, the integration of spaCy into our methodology exemplifies the tool's adeptness in linguistic feature extraction and contributes to the broader research aim of elucidating the explainability of neural model decisions.

5.2. BERT Explainability

Statistical tests utilizing SHAP values indicate that BERT models can distinguish between speeches and interviews based on linguistic features. The significant p-values for features like nominalization frequency, discourse marker frequency, personal pronoun frequency, and interjection frequency in BERT1 demonstrate that these features are part of the model's classification decisions. The differences in means and standard deviations in the t-tests between speeches and interviews for these features (Table 6) confirm that idea. This relationship indicates that BERT's classification decisions can be partially explained by its sensitivity to these linguistic features, answering RQ1 positively.

We anticipated that removing named entities would make the model focus on defined linguistic features, potentially simplifying interpretation. However, our findings show that the model may have adopted more abstract or distributed linguistic cues, which are not as easily captured by our feature set, thereby complicating interpretability. This observation is crucial for answering RQ2, where the diminished significance of most features in BERT2 could point to a reduced sensitivity to the specific linguistic characteristics that differentiate speeches from interviews, implying that the BERT2 model may be leveraging other aspects of the data or relying on a more generalized understanding of the text to make its classifications. This finding highlights the complexity of BERT's decision-making processes and the limitations of current explanatory tools in capturing the entirety of these processes.

Therefore, RQ2 is partially answered since the performance metrics of BERT2 prove that the model relies less on thematic data but not on the studied linguistic features, as we expected; all that without detriment of its performance. The following are explanations that may contribute to our understanding of BERT2, performing with high accuracy, as shown in Table 5:

- Despite anonymizing identifiable entities, BERT2 might still be leveraging the residual semantic context that is not removed by NER anonymization. This context includes thematic elements, narrative flow, and the abstract representation of concepts that remain encoded in the text, allowing the model to distinguish between speeches and interviews based on the thematic undercurrents of the discourse.

- The BERT model's architecture enables it to capture intricate interactions between linguistic features that transcend simple lexical frequencies or easily identifiable linguistic rules. These interactions may involve a nuanced understanding of language, including how different linguistic elements coalesce to convey meaning, tone, or stylistic nuances specific to speeches or interviews.

- The BERT model's embeddings and hidden layers encode information in high-dimensional spaces, abstractly representing text in a way that does not directly align with human linguistic

concepts. Consequently, the BERT model may base its decisions on patterns within these spaces that are obscure to SHAP analyses, suggesting a layer of complexity in model decision-making that extends beyond conventional linguistic or semantic analysis.

One possible explanation for the previous ideas is the BERT's increased reliance on subwords after the limitation that NER anonymization brings to the model. Analyzing the word length (token length) of SHAP-attributed features (tokens) before and after NER anonymization, we observed a shift in the model's focus from relying on potentially identifying longer tokens ("*whole*" words) to shorter tokens (subwords). This observation aligns with our earlier decision to prune word length from LogRegFeat, as BERT implicitly captures word length through subword tokenization. When named entities were removed, the model lost access to certain content-heavy words, leading to increased reliance on fragmented subword tokens. While a t-test showed no statistically significant difference in token length between conditions ($p = 0.325$), Figure 9 illustrates a clear tendency towards shorter tokens after the NER anonymization treatment. This observation highlights how NER anonymization influences the linguistic features SHAP identifies as significant, reducing the average word length within both classes and reinforcing BERT's reliance on subword-level representations.

To clarify the impact of NER anonymization on our analysis, we examined the tokens that exhibited the highest SHAP values in each class post-anonymization. In the speech class, these tokens included fragments like "*contra*", "*ceptive*", "*ita*", "*oche*", and "*shi*", while in the interview class, they were "*osing*", "*du*", "*ita*", "*dly*", and "*ep*". These fragments appear to be suffixes or subwords, indicating a trend towards shorter word lengths in BERT2. This observation shows how NER anonymization can influence the linguistic features SHAP identifies as significant by reducing the average word length within both classes.

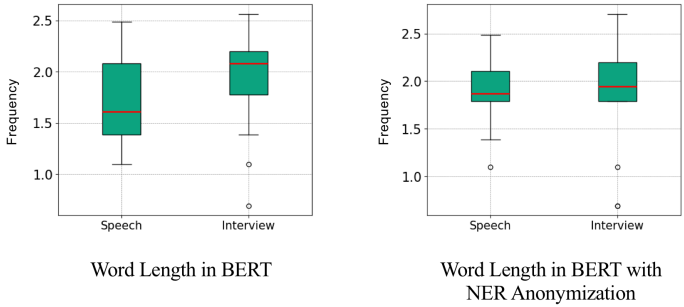

Figure 9: Comparison of Token (Word) Length of SHAP attributed features between BERT models with and without Anonymization for Speech and Interview Classification.

6. Conclusion

Our investigation shed light on the differences in linguistic structures between speeches and interviews, showcasing their inherent interdependence and the need to embrace these complexities to develop methods capable of capturing them. Statistical analyses, supported by SHAP values, revealed that nominalization frequency, discourse marker frequency, personal pronoun frequency, and interjection frequency significantly influence the BERT model's decision-making process in the non-anonymized dataset, TextDataset. Consequently, our study contributes to expanding knowledge in explainable NLP and AI by providing empirical evidence of the role of concrete linguistic features in the classification decisions of advanced NLP models. This insight increases the BERT model's transparency and solidifies our understanding of how NLP processes the nuanced subtleties of political language.

Regarding the toolbox used in this research, while spaCy performed with a high degree of accuracy in identifying linguistic features in the text by capturing the complexities of political language, its operationalization at the semantic and pragmatic levels remains a challenge for this generation of pre-trained NLP models. The tool's ability to precisely identify interjections is noteworthy, as these subtle differences are crucial in the studied classification task. The precise detection of linguistic structures in political communication without the hassle of training ML models contributed enormously to the explainability of the NLP models and the comprehension of linguistic phenomena.

As noted earlier, the classification task between speeches and interviews is relatively straightforward for machines and humans, compared to many real-world classification problems where discourse characteristics are more ambiguous—something reflected in both the IAA ($\kappa = 0.956$) and model accuracy (BERT1: $Acc = 0.981$; BERT2: $Acc = 0.974$). Therefore, we do not think the methodology is directly transferrable to more complex or subtle classification tasks, like sarcasm detection or speaker stance classification, where features operate at linguistics' semantic or pragmatic level. Nonetheless, the explainability framework we employ, especially the combination of LRBM with SHAP-based analysis, could serve as a modular approach to enhance transparency even in more challenging NLP tasks.

To some extent, our research could be applied to different domains where the discrimination of speeches and interviews may be necessary, but 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. This specificity aids in minimizing external variability, but it also introduces constraints in terms of generalizability to political discourse in other domains or political systems. On the other hand, systematic curation and preprocessing while building the datasets were important steps in reducing semantic and context bias in the models. Also, our inability to distinguish between planned (probably reading a script) and spontaneous speeches in the corpus-building stage introduces a significant blind spot, where the nature of the discourse can significantly affect its content and structure, potentially influencing the model's ability to classify and analyze the data accurately.

The reliance in our study mostly on specific morphosyntactic features partially overlooks semantic or pragmatic aspects of political discourse, such as sentiment, stance, or thematic content, which could also be critical in differentiating speeches from interviews. Although morphology

and syntax complexity may be important in the classification, the context in which words are used (even beyond anonymization); and the presence of specific rhetorical devices could add an even deeper layer of understanding to the analysis. However, it is noteworthy that, as seen in LogRegFeat results (Table 4), five linguistic features were enough to achieve high performance in distinguishing speeches and interviews. Other limitations of this research include (1) the gender imbalance in the dataset, with 98.5% of speakers being male; (2) the restriction to American English political discourse, which may limit cross-cultural generalizability; and (3) the SHAP framework's limitation to token-level explanations, which constrains its ability to capture phrase-level linguistic features such as passive voice or sentence complexity.

While SHAP demonstrated utility in elucidating aspects of BERT models' decision-making processes, in the context of utilizing NER anonymization, it unveils a paradox of NLP/AI performance and explainability: the intriguing capacity of BERT models to maintain classification accuracy by relying on unrecognizable linguistic cues. We observed that as BERT adapts to anonymization by extracting meaning from obscured linguistic cues, the complexity of its decision-making processes increases, challenging the current capabilities of tools like SHAP to provide transparent explanations. Consequently, anonymization techniques like NER might enhance privacy and reduce bias by removing identifiable information without sacrificing the BERT model's performance. Therefore, NER anonymization introduces an additional layer of complexity to NLP/AI explainability, spotlighting the need for further studies in this area.

To confirm these results, preregistered replications using independent corpora and alternative annotation pipelines would be valuable, ideally with designs that match speakers and topics across interviews and speeches and include mixed-effects models to account for speaker/outlet variation.

A. Lexicons

Lexicons used to operationalize the studied linguistic features:

1. Personal Pronouns

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

2. Nominalization (Suffixes)

-tion, -ment, -ness, -ity, -age, -ance, -ence, -hood, -ship, -ty, -cy, -al, -ure, -er, -ing, -sion, -ation, -ibility, -ana, -acy, -ama, -ant, -dom, -edge, -ee, -eer, -ery, -ese, -ess, -ette, -fest, -ful, -iac, -ian, -ie, -ion, -ism, -ist, -ite, -itude, -ium, -let, -ling, -man, -woman, -mania, -or, -th, -tude.

3. Discourse Markers

accordingly, actually, after, after all, afterward, afterwards, all in all, also, although, always, anyway, apparently, as, as a matter of fact, as a result, as if, as long as, as soon as, as though, as well as, assuming that, at first, at last, at least, at present, at the same time, basically, because, before, being that, besides, but, by the time, by the way, chiefly, clearly, commonly, consequently, considering that, despite, due to, during, either, especially, essentially, even if, even so, even supposing, even though, eventually, every, evidently, except, except for, except that, exclusively, finally, first, first of all, for example, for instance, for one thing, for the most part, for the time being, for this purpose, for this reason, formerly, forthwith, fortunately, frankly, frequently, further, furthermore, generally, given that, granting that, hence, henceforth, honestly, however, if only, immediately, in addition, in any case, in brief, in case, in conclusion, in contrast, in fact, in general, in my opinion, in order that, in order to, in other words, in particular, in short, in spite of, in sum, in summary, in that case, in the beginning, in the end, in the first place, in the meantime, in the meanwhile, in the same way, in the second place, in the third place, in this case, in truth, in view of, incidentally, including, indeed, individually, initially, instantly, instead, interestingly, just, largely, last, lastly, lately, later, lest, like, likewise, mainly, markedly, meanwhile, merely, moreover, most, most importantly, mostly, much, namely, naturally, neither, never, nevertheless, next, nonetheless, nor, normally, not just, not only, not to mention, notably, notwithstanding, now, now that, nowadays, obviously, occasionally, of course, often, on balance, on condition that, on the contrary, on the other hand, on the whole, once, only, ordinarily, originally, otherwise, overall, particularly, permanently, personally, plainly, plus, presently, presumably, previously, primarily, privately, probably, promptly, properly, provided that, publicly, quickly, rarely, rather, readily, really, recently, regardless, regularly, relatively, remarkably, respectively, secondly, seeing that, seemingly, seldom, separately, seriously, shortly, significantly, similarly, simply, since, slightly, slowly, so, so far, so long as, so that, sometimes, soon, specifically, still, straightaway, strangely, strongly, subsequently, successfully, such as, suddenly, supposedly, supposing, surely, surprisingly, technically, temporarily, that is why, then, thereafter, thereby, therefore, thereupon, these days, thirdly, though, thus, till, to begin with, to conclude, to date, to illustrate, to sum up, to tell the

truth, to that end, to this end, typically, ultimately, undoubtedly, unfortunately, unless, unlike, until, until now, up to, up to now, up to the present time, usually, utterly, virtually, well, what is more, whatever, when, whenever, whereas, whereby, whereupon, wherever, whether, whichever, while, whilst, whoever, with this in mind, with this intention, yet.

4. Interjections

ah, alright, amen, anyway, augh, aye, boy, boom, bye, congratulations, darn, ehh, er, fine, foul, gee, god, goodby, goodbye, goodness, gosh, gracious, gross, heavens, heck, hello, hey, hi, holy, hooray, huh, kid, like, man, nah, nay, naw, no, nope, now, ok, okay, oh, oops, please, right, say, see, shit, sorry, sure, thank, thanks, uh, um, welcome, well, whoa, whoop, whoops, wow, yea, yeah, yep, yes.

B. Annotation Guidelines of Dataset 1: Speeches and Interviews

Author: Juan-Francisco Reyes

NOTE: THIS IS AN ADAPTED VERSION OF THE ORIGINAL ANNOTATION PROCEDURE, RETAINING CONTENT RELEVANT TO THIS PAPER.

This document delineates the process of cleaning, anonymizing, and annotating a dataset for training a text classification model that distinctly identifies two discourse text classes: *speech and interview*. Participants will generate subsets of the "*Corpus 1*" repository, equally split between instances of speeches and interviews. Utilizing an annotation tool, participants will process texts into datapoints to build a dataset.

Task

Participants are to construct a dataset comprising audio/transcribed political discourses, with equal representation of speech and interview examples.

Selection Criteria

The selection criteria we follow to add discourse texts to the dataset include:

- *Domain*: The texts pertain solely to the US political sphere, with American English as the language medium.
- *Representativity*: Selected texts represent perfect examples of speeches and interviews, without any elements that may distort the discourse class.
- *Interaction*: To prevent model misinterpretation of the speech class, speeches with high speaker-audience interactions should be skipped or adjusted to maintain the discourse class integrity. For instance, processing speeches by individuals like Donald Trump, known for engaging audiences with rhetorical questions, necessitates caution.

Step 1: Access the Annotation Tool

Enter the Annotation Tool web application, logging in with credentials supplied by the lecturer.

Step 2: Load Text Editor

Access the text editor, select one of the assigned datapoints you intend to annotate, and recognize the following areas and elements (Figure 10):

- *Text area*: The text editing area, accompanied by a Word Counter display below.
- *Sidebar*: The area that houses the discourse text Title, Source Link, Class Selector, Save Button, and Timestamp Remover tool.

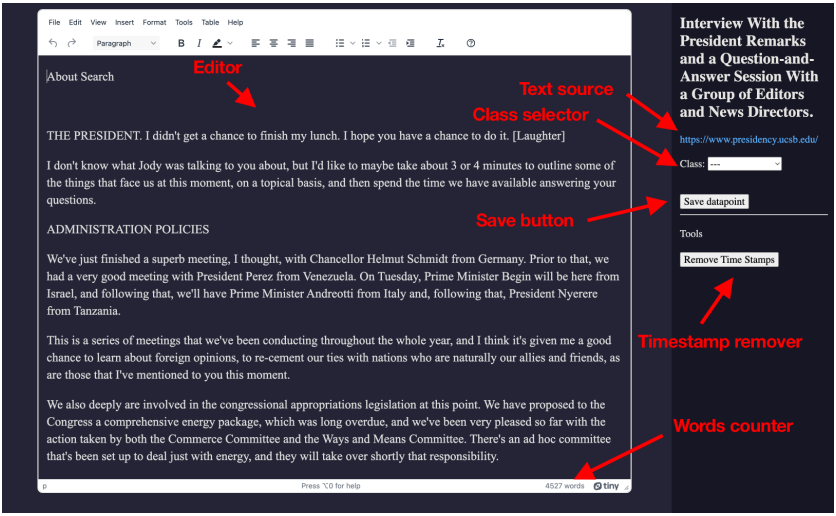

Figure 10: Annotation tool interface.

Step 3: Remove Timestamps

The Timestamp Remover tool automatically removes the time stamps from transcriptions of most interviews and speeches. For instance, the original text “Donald Trump (00:02):”, after applying the remover, results in “Donald Trump:”. This tool matches different patterns, like “Donald Trump [00:02]:” or “Donald Trump (01:15:02):”. If the Timestamp Remover tool does not match the timestamp patterns in an assigned document, you will have to remove them manually.

Step 4: Check Word Counter

Use the Word Counter to check the text length, and if the text discourse does not have at least 450 words, the document must be "ignored" (skipped). If an excessive number of documents are ignored, and your assigned batch of text discourses is consumed, request additional text discourses from the lecturer. Notice that counting the minimal number of words is a permanent task during the whole text cleaning and processing.

Step 5: Classify and Label the Text

Observe the document and label it either with the class "SPEECH" or "INTERVIEW" by using the Class Selector. If necessary, refer to the original text (using the Source Link) to understand the nature of the discourse. Some discourse texts may pose classification challenges, but in general, it is very easy to discriminate them by simple observation. Assess the ease of classifying the document; if it is troublesome, skip it. Occasionally, minor tweaks—like removing brief interactions from one participant—maintain a text discourse in the speech class. On the contrary, if a second speaker's involvement is prevalent, consider the document in the interview class, ignore it, or "tweak" it. Recall that we must include perfect examples of each class in the dataset, and it is allowed to make minor editions (mostly minor removals) to make the class features more salient. Notice that by using the Class Selector, you are labeling (annotating) the text and defining how the NLP model will behave.

Hybrid discourses: Sometimes, a speech (monologic discourse) ends with a conference (dialogic discourse). If, after removing one of both segments, the remainder of the text is a good example of one of both classes, then removing large parts of the original text is valid. Check the Word Counter to evaluate the removal of large text segments.

Step 6: Remove Surrounding Text

Surrounding text refers to any content not integral to the primary discourse. The discourse should commence where the main speaker(s) distinctly begin their participation. The onset may include greetings or not, and similarly, it may or may not conclude with farewells. Surrounding text are introductory remarks by other speakers prior to a speech or interview or introductions written by editors. Surrounding text acts as noise and its elimination from the datapoint is essential for a clearer analysis.

Step 7: Identify Participants in Interviews

In speeches, only one participant is involved, but in interviews, a minimum of two participants is required, and a maximum of four is allowed. Interviews typically comprise at least one "*Interviewer*" and one "*Interviewee*". For example, scenarios with two individuals conducting the interview (interviewers) or being interviewed (interviewees) are acceptable. The configuration of two interviewers and two interviewees represents the maximum number of participants for interviews. Given the complexity of participant/speaker variance, extra caution is advised when annotating interviews.

Step 8: Anonymize Speaker(s)

Anonymizing speeches is simpler compared to interviews, as the first typically features a single-speaker label, while the latter contains multiple. In interviews, careful attention to speaker labels and cross-references is essential. Cross references in interviews are mentions of one participant by another participant—see the applied example below for reference. For this text classification use case, enhancing the saliency of linguistic features for the model training is crucial, and the presence of (person) names could mislead the model into classifying texts based on person names, organizations, or locations. Therefore, anonymization strips contextual information, enabling the model to learn from the linguistic structures rather than context.

The task of anonymization applies solely to speakers engaged in the discourse, where their labels (e.g., "DONALD TRUMP:", "DT:", or "A." [from "*Answer*"]) or cross-references are replaced with *placeholders*. For example, in an interview featuring Donald Trump as a speaker, his name should be anonymized; however, if another participant merely mentions him, no anonymization is required. *Named entities* mentioned in the discourse (e.g., "*White House*", "*N.A.T.O.*", or "*Germany*") should not be anonymized. Utilize the "*Find and Replace*" feature (Ctrl + F) in the editor to expedite this step.

Step 9: Remove Noisy Text

As you have learned in the Hugging Face's NLP Course as part of this seminar, deep learning models learn by "seeing" good examples of the classes we want to predict; therefore, removing noise is another important step. Noisy text are notations on, for instance, applauses ("[*applause*]"), cheers ("[*crowd cheering*]"), onomatopoeic expressions ("*Wooo!*", "*Boo!*"), etc., that have been labeled or transcribed in the text. Those patterns can be found in speeches and interviews and must be removed to avoid the model paying attention to them, forcing it to focus on discourse structures. Again, the "Find and Replace" feature is also very useful in this step.

Step 10: Save the Datapoint

Use the Save Button to save the datapoint.

Demo: Annotating an Interview (Video)

Watch the provided demo video `https://www.youtube.com/watch?v=9fg6bwicFhs` for guidance on annotating an interview.

Applied Example

Open the provided interview and notice the omission of irrelevant, noisy, and contextual information at the outset, enhancing the clarity of the beginning of the discourse. The interview commences with "*With that, we're open to questions, . . .* " indicating the transition to a dialogic discourse. Note the labeling of the interviewer initially as "Q." (from "*Question*") later anonymized to "[INTERVIEWER]". Additionally, observe the anonymization of "MR. FROMAN" (Michael

B. Froman) and "MR. STERN" (Todd Stern) to "[INTERVIEWEE1]" and "[INTERVIEWEE2]", respectively.

Excerpts of the original text:

> "Q: On trade, Mike, would you say this promise to restart Doha in 2010, is that a reflection of the economic and political realities of the global crisis? Is that something that the President – President Obama supports, or was there any effort by anyone to try to see if it could get on track earlier than that?"
>
> Q: Thanks. Mike or Todd, could you give us some color about the President's role at the MEF meeting today – whether he – where he (inaudible), what actual effect he had on the final result? And I'll repeat a question that I asked at the earlier briefing, which was, the President had said he wants the United States to show leadership on climate change. Did he achieve that (inaudible)?
>
> MR. FROMAN: I think he certainly achieved that. I think that there's wide recognition and wide appreciation, actually, of the role of the United States and the change that the President has brought in U.S. policy on this issue, which has been more dramatic perhaps than in any other area.

Excerpts of the original anonymized text:

> "[INTERVIEWER]: On trade, [INTERVIEWEE1], would you say this promise to restart Doha in 2010, is that a reflection of the economic and political realities of the global crisis? Is that something that the President – President Obama supports, or was there any effort by anyone to try to see if it could get on track earlier than that?"
>
> [INTERVIEWER]: Thanks. [INTERVIEWEE1] or [INTERVIEWEE2], could you give us some color about the President's role at the MEF meeting today – whether he – where he (inaudible), what actual effect he had on the final result? And I'll repeat a question that I asked at the earlier briefing, which was, the President had said he wants the United States to show leadership on climate change. Did he achieve that (inaudible)?
>
> [INTERVIEWEE1]: I think he certainly achieved that. I think that there's wide recognition and wide appreciation, actually, of the role of the United States and the change that the President has brought in U.S. policy on this issue, which has been more dramatic perhaps than in any other area.

Peer Review, Inter-Annotator Agreement (IAA)

The objective is to ensure a high degree of agreement among annotators in classifying (into speeches and interviews), cleaning, and anonymizing political discourses. This IAA assessment will aim to verify the consistency and reliability of the annotations, which is crucial for the subsequent analysis involving the model training.

Each discourse text will be annotated by at least two different annotators to ensure redundancy. The second annotator—the reviewer—will annotate the agreement assessment based on the following six facets:

1. *Correct Class*: If the classification is correct.
2. *Cleaned Timestamps*: If timestamps were removed.
3. *Cleaned Surrounding*: If the surrounding text was removed.
4. *Cleaned Noise*: If applause, cheers, and other interactions with the public were removed.
5. *Anonymized Speaker*: If anonymization of speakers is correct.
6. *Anonymized Speaker Cross-Reference*: If the anonymization of the speakers cross-references is correct.

In the shared *Google Spreadsheet*, add a cross, "X", whenever a discrepancy is found. Leave the cell in blank if no discrepancy was found. Immediately after, make corrections to the datapoint, and save it.

If a discrepancy is found, it should be discussed and resolved in the weekly sessions. Attend the review meetings to discuss disagreements and clarify any ambiguities in the guidelines, helping to improve the annotation quality moving forward.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019)*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Akpatsa, S. K., Li, X., & Lei, H. (2021). A survey and future perspectives of hybrid deep learning models for text classification. *Artificial Intelligence and Security*, 358–369. Springer. https://doi.org/10.1007/978-3-030-78609-0_31

Ameka, F. K. (1992). Interjections: The universal yet neglected part of speech. *Journal of Pragmatics*, 18(2–3), 101–118. https://doi.org/10.1016/0378-2166(92)90048-G

Amelia, P., Sinar, T., & Zein, T. (2020). Lexical density and grammatical intricacy in debaters' speeches. *Languaje Literacy: Journal of Linguistics, Literature, and Language Teaching*, 4(1), 168–184. https://doi.org/10.30743/ll.v4i1.2519

Austin, J. L. (1962). *How to Do Things with Words*. Clarendon Press.

Bakhtin, M. M. (1981). *The Dialogic Imagination: Four Essays*. University of Texas Press.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, 15(2), 133–163. https://doi.org/10.1080/01638539209544806

Biber, D. & Finegan, E. (1994). *Sociolinguistic Perspectives on Register*. Oxford University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education.

Billig, M. (2008). The language of critical discourse analysis: the case of nominalization. *Discourse & Society*, 19(6), 783–800. https://doi.org/10.1177/0957926508095894

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, 538(7623), 20–23. https://doi.org/10.1038/538020a

Chilton, P. (2004). *Analyzing Political Discourse: Theory and Practice*. Routledge.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. (2019). What does BERT look at? an analysis of BERT's attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276–286. https://doi.org/10.18653/v1/W19-4828

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viegas, F., & Wattenberg, M. (2019). Visualizing and measuring the geometry of BERT. *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS'19)*, 8594–8603. https://doi.org/10.48550/arXiv.1906.02715

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *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), volume 1, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dingemanse, M. (2023). Interjections. *Oxford Handbook of Word Classes*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198852889.001.0001

Dingemanse, M. (2024). Interjections at the heart of language. *Annual Review of Linguistics*, 10, 257–277. https://doi.org/10.1146/annurev-linguistics-031422-124743

Du Bois, J. W. (2007). The stance triangle. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, 139–182. John Benjamins Publishing Company. https://doi.org/10.1075/pbns.164.07du

Du Bois, J. W. (2014). Towards a dialogic syntax. *Cognitive Linguistics*, 25(3), 359–410. https://doi.org/10.1515/cog-2014-0024

Fairclough, N. (2001). *Language and power* (2nd ed.). Longman.

Fowler, R., Hodge, B., Kress, G., & Trew, T. (1979). *Language and Control*. Routledge & Kegan Paul.

Grzybek, P., Stadlober, E., & Kelih, E. (2006). The relationship of word length and sentence length: The inter-textual perspective. *Advances in Data Analysis*, 611–618. Springer. https://doi.org/10.1007/978-3-540-70981-7_70

Gumperz, J. J. (1982). *Discourse Strategies*. Cambridge University Press.

Halliday, M. A. K. (1994). *An Introduction to Functional Grammar*. Edward Arnold.

Heeman, P., Byron, D., & Allen, J. (1998). Identifying discourse markers in spoken dialog. *Proceedings of the Applying Machine Learning to Discourse Processing*, 44–51. https://doi.org/10.48550/arXiv.cmp-lg/9801002

Hirst, G. (2001). Longman grammar of spoken and written english. *Computational Linguistics*, 27(1), 132–139. https://doi.org/10.1162/089120101300346831

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in python*. https://doi.org/10.5281/zenodo.1212303

Htut, P., Phang, J., Bordia, S., & Bowman, S. (2019). *Do attention heads in bert track syntactic dependencies?* https://doi.org/10.48550/arXiv.1911.12246

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

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. Continuum.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. https://doi.org/10.18653/v1/P19-1356

Jumelet, J. & Zuidema, W. (2023). Feature interactions reveal linguistic structure in language models. *Findings of the Association for Computational Linguistics: ACL 2023*, 8697–8712. https://doi.org/10.18653/v1/2023.findings-acl.554

Jurafsky, D. & Martin, J. H. (2025). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd ed.). Online manuscript. https://web.stanford.edu/~jurafsky/

slp3/

Kashiha, H. (2021). Stance-taking across monologic and dialogic modes of academic speech. *Southern African Linguistics and Applied Language Studies*, 39(4), 352–362. https://doi.org/10.2989/16073614.2021.1964371

Kazemian, B. & Hashemi, S. (2014). Nominalizations in scientific and political genres: A systemic functional linguistics perspective. *International Journal of Humanities and Social Sciences*, 3(2), 211–228. https://ssrn.com/abstract=2514388

Kitagawa, C. & Lehrer, A. (1990). Impersonal uses of personal pronouns. *Journal of Pragmatics*, 14(5), 739–759. https://doi.org/10.1016/0378-2166(90)90004-W

Koplenig, A. (2019). Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. *Royal Society Open Science*, 6(2). https://doi.org/10.1098/rsos.181274

Larsson, T. & Kaatari, H. (2020). Syntactic complexity across registers: Investigating (in)formality in second-language writing. *Journal of English for Academic Purposes*, 45. https://doi.org/10.1016/j.jeap.2020.100850

Lee, K., Dobbins, N. J., McInnes, B., Yetisgen, M., & Uzuner, Ö. (2021). Transferability of neural network clinical deidentification systems. *Journal of the American Medical Informatics Association*, 28(12), 2661–2669. https://doi.org/10.1093/jamia/ocab207

Lei, J., Rahman, T., Shafik, R., Wheeldon, A., Yakovlev, A., Granmo, O., Kawsar, F., & Mathur, A. (2021). Low-power audio keyword spotting using tsetlin machines. *Journal of Low Power Electronics and Applications*, 11(2), 18. https://doi.org/10.3390/jlpea11020018

Li, B. (2022). *Integrating Linguistic Theory and Neural Language Models*. University of Toronto. https://doi.org/10.48550/arXiv.2207.09643

Li, Z., Zhou, Q., Li, C., Xu, K., & Cao, Y. (2020). Improving BERT with syntax-aware local attention. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 645–653. https://doi.org/10.18653/v1/2021.findings-acl.57

Liu, M. (2022). A corpus-based study on the usage of passive voice in english political speeches on the guidance of text typology. *The Frontiers of Society, Science and Technology*, 4(1). https://doi.org/10.25236/fsst.2022.040113

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. *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)*, 1073–1094. https://doi.org/10.18653/v1/N19-1112

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30 of *NIPS'17*, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

Mauranen, A. (2023). *Reflexively speaking: Metadiscourse in English as a lingua Franca*. De Gruyter. https://doi.org/10.1515/9783110295498

McCarthy, M. & Carter, R. (1995). Spoken grammar: What is it and how can we teach it? *ELT Journal*, 49(3), 207–218. https://doi.org/10.1093/elt/49.3.207

Mendhakar, A. (2022). Linguistic profiling of text genres: An exploration of fictional vs. non-fictional texts. *Information*, 13(8), 357. https://doi.org/10.3390/info13080357

Michel, P., Levy, O., & Neubig, G. (2019). Are sixteen heads really better than one? *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 14037–14047. https://doi.org/10.48550/arXiv.1905.10650

Mosca, E., Demirtürk, D., Mülln, L., Raffagnato, F., & Groh, G. (2022). GrammarSHAP: An efficient model-agnostic and structure-aware NLP explainer. *Proceedings of the First Workshop on Learning with Natural Language Supervision*, 10–16. https://doi.org/10.18653/v1/2022.lnls-1.2

Nikolaidis, K., Kristiansen, S., Plagemann, T., Goebel, V., Liestøl, K., Kankanhalli, M., Traaen, G. M., Øverland, B., Akre, H., Aakerøy, L., & Steinshamn, S. (2021). Learning realistic patterns from visually unrealistic stimuli: Generalization and data anonymization. *Journal of Artificial Intelligence Research*, 72, 1163–1214. https://doi.org/10.1613/jair.1.13252

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. *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

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

Raskin, A. & Harris, T. (2023). *The A.I. dilemma*. https://singjupost.com/discussion-the-a-i-dilemma-march-9-2023-transcript/

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

Reyes, J. F. (2026). *Hugging Face models and datasets*. https://huggingface.co/pacoreyes

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8, 842–866. https://doi.org/10.1162/tacl_a_00349

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn taking in conversation. *Language*, 50, 696–735. https://doi.org/10.2307/412243

Seabold, S. & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python.

Proceedings of the 9th Python in Science Conference, 57–61. https://doi.org/10.25080/Majora-92bf1922-011

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Tannen, D. (1981). *Conversational Style: Analyzing Talk Among Friends*. Ablex Pub. Corp.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601. https://doi.org/10.18653/v1/P19-1452

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

Tottie, G. (1991). Conversational style in british and american english: The case of interjections. *Journal of Pragmatics*, 15(1), 13–28.

Van Dijk, T. A. (1998). *Ideology: A Multidisciplinary Approach*. Sage Publications.

Vanni, L., Corneli, M., Mayaffre, D., & Precioso, F. (2023). From text saliency to linguistic objects: learning linguistic interpretable markers with a multi-channels convolutional architecture. *Corpus*. https://doi.org/10.4000/corpus.7667

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

Wells, G. (2006). Monologic and dialogic discourses as mediators of education. *Research in the Teaching of English*, 41(2), 168–175. https://www.jstor.org/stable/40039099

Xiaomao, X., Xudong, Z., & Yuanfang, W. (2019). A comparison of feature selection methodology for solving classification problems in finance. *Journal of Physics: Conference Series*, 1284. https://doi.org/10.1088/1742-6596/1284/1/012026

Yin, K. & Neubig, G. (2022). Interpreting language models with contrastive explanations. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 184–198. https://doi.org/10.18653/v1/2022.emnlp-main.14

Zafar, M., Donini, M., Slack, D., Archambeau, C., Das, S., & Kenthapadi, K. (2021a). On the lack of robust interpretability of neural text classifiers. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3730–3740. https://doi.org/10.18653/v1/2021.findings-acl.327

Zafar, M., Schmidt, P., Donini, M., Archambeau, C., Biessmann, F., Das, S., & Kenthapadi, K. (2021b). *More than words: Towards better quality interpretations of text classifiers*. https://doi.org/10.48550/arXiv.2112.12444

Zare, J. & Tavakoli, M. (2016). The use of personal metadiscourse over monologic and dialogic modes of academic speech. *Discourse Processes*, 54(2), 163–175. https://doi.org/10.1080/0163853X.2015.1116342

Zhang, Y., Zhang, P., & Yan, Y. (2019). Tailoring an interpretable neural language model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(7), 1164–1178. https://doi.org/10.1109/TASLP.2019.2913087

Zhao, W., Joshi, T., Nair, V., & Sudjianto, A. (2020). *SHAP values for explaining CNN-based text classification models*. https://doi.org/10.48550/arXiv.2008.11825

Correspondence

Juan-Francisco Reyes

Brandenburgische Technische Universität Cottbus-Senftenberg
Fakultät 1 – Institut für Informatik
Cottbus, Germany
pacoreyesp@gmail.com